JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

Breast Cancer Diagnosis by Exploiting the Permutations of Principal Components by Ensemble Classification

Aimen Sikander[1], and Iqbal Murtza[1*]

[1]Faculty of Computing & AI, Air University, Islamabad, Pakistan.
[*]Corresponding Author: Iqbal Murtza. Email: iqbal.murtza@mail.au.edu.pk

Received: January 01, 2024 **Accepted:** February 26, 2024 **Published:** March 01, 2024

Abstract: In many breast cancer computer-aided diagnosis problems with larger feature dimensions and fewer feature instances, the classification does not get optimal training. This is because a decision boundary is represented by the number of parameters directly proportional to the feature dimensions. Since the optimal training of such high-dimensional features requires a large training set. Unluckily, if the training set is not sufficiently large to generate good n/l ratio, the training results in an ineffective and inefficient classification model. To resolve the problem of large dimensions, the conventional employment of feature reduction techniques results in efficient training however it yields the degraded classification performance. In this paper, we consider this problem to have effective and efficient training in large dimensional datasets when the available dataset is not sufficiently large. For this purpose, we hybridize principal component analysis with ensemble classification. For this, different combinations of principal dimensions have been determined by the concept of power sets in mathematics. A dedicated base learner then exploits each principal dimension combination. Then, all these base learners are combined to construct a hybrid ensemble principal component analysis-based classifier, Ens-PCA. The proposed Ens-PCA technique is tested using Wisconsin diagnostic breast cancer (WDBC) data set and the results show its outperformance as compared to the contemporary principal component analysis and ensemble classification techniques.

Keywords: Machine learning; Ensemble classification; Principal component analysis; set theory; Data sampling.

1. Introduction

In recent decades, the statistics on breast cancer are alarming. Each year around one million ladies are identified with breast cancer globally. In 2020 around 2300 thousand new cases were identified with women breast cancer. Unluckily, breast cancer has become the most frequently detected cancer, with 2.3 million new cases [1, 2]. According to recent insights into Breast cancer, Pakistan has the highest rate of breast cancer in Asia. Every year, 90000 new patients are diagnosed with breast cancer, and almost 40000 patients die [3]. Breast cancer can be cured sufficiently with up to an 81% survival rate if diagnosed in the initial state. Most women are later diagnosed with cancer because it's asymptomatic and spreads to other organs. Breast cancer originates primarily in adipose tissues, connective tissues, ducts, and lobules of the breast [4]. In this regard, accurate diagnosis of cancer is very critical. Recently, the deployment of machine learning techniques for computer-aided cancer diagnosis is getting the attention of the research community [5, 6].

In many cancer computer-aided diagnosis problems, the task of classification is for precise and accurate decisions. However, in many problems, the presence of larger dimensional feature space affects the classification and degrades its performance. This is because, the higher dimensional feature space has associated problems such as the curse of dimensionality, inefficient training, and ineffective testing [8-10]. This is because, the classification boundary is represented by the number of parameters proportional to the

number of feature dimensions. For example, a single linear discriminating in n dimensional feature space is represented by $2n$ parameters, whereas, second order single discriminating boundary needs $3n$ parameters.

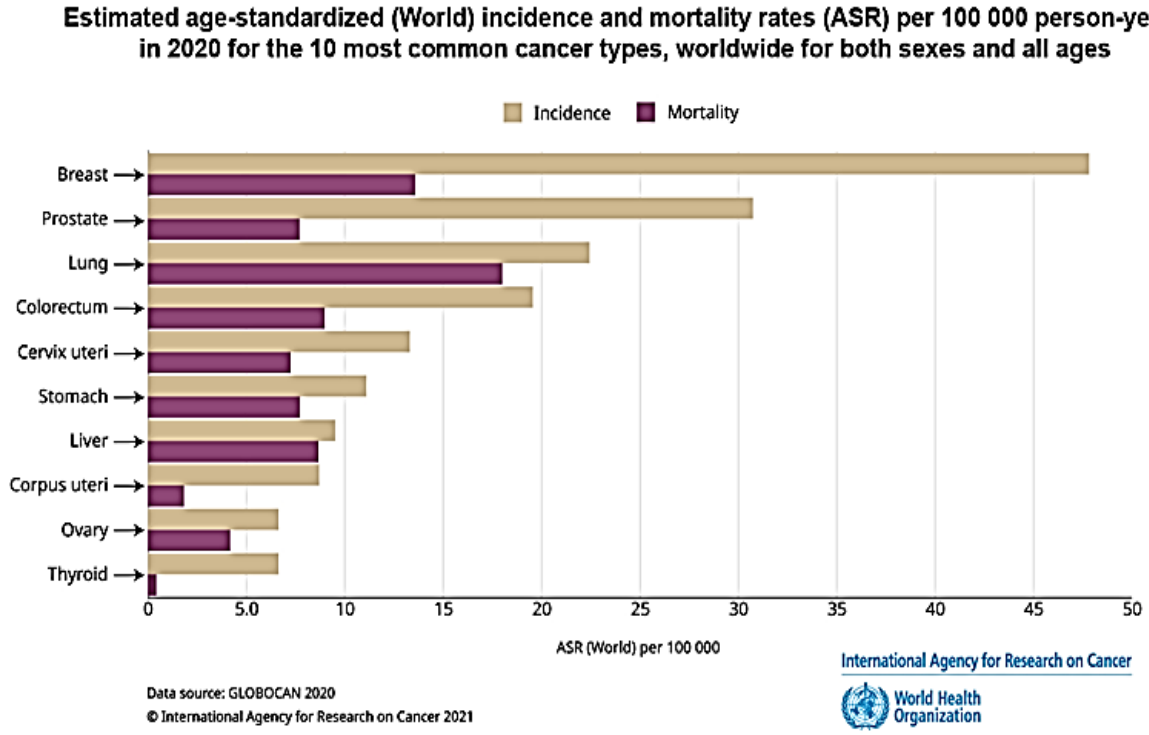


Figure 1. Worldwide statistics of various cancers reported by the international agency for research on cancer, world health organization (WHO), showing the breast cancer is the most frequent [7]

Because of the above proportional parameter requirements in the classification model, the training phase of the most classification techniques uses gradient descent to optimize the choice of these parameters. Unluckily, the efficiency of such optimization techniques degrades as the number of parameters increases which results in slow training [8]. In addition to this, the chance of finding the global optimal point in this parameter space becomes less probable as well which results in an immature trained classification model resulting in a downgraded classification of samples in testing phase [11, 12]. To address this curse of dimensionality, the employment of feature reduction techniques, although gained substantial attention of the research community but, it results in downgraded classification performance. The reason of this downgraded performance is because of the information loss which is although partial but it plays a role in the degradation [13]. In this paper we consider this problem to boost the classification performance affected by feature reduction techniques.

Principal component analysis (PCA) is a technique for dimensionality reduction, which can potentially improve the accuracy of a model by reducing the number of input features. It is a classical statistical method for transforming attributes of a dataset into a new set of uncorrelated attributes called principal components (PCs). In this way, it is capable extracting compact information from high dimensional spectra while maximally retaining as much of the variability of the dataset as possible [9]. This process is accomplished by eigenvector decomposition followed by neglecting the unimportant directions in which sample variances are insignificant such that the number of these directions approximates the dimensionality of the whole sample set. Alongside these strengths, unluckily, PCA has some limitations as follows:

1. It is sensitive to the scale of the input features. It is generally recommended to standardize the data before applying PCA [8].
2. It is a linear technique and may not be appropriate for data that is highly nonlinear [9].
3. It is a linear technique and does not capture non-linear interactions between variables [14].
4. It is an unsupervised technique and does not take into account any labels or class information.
5. It is not robust to outliers, which can have a disproportionate effect on the principal components [15].

6. It only considers the most important k principal components, which may not be enough to capture all the information in the data.
7. It can be affected by correlated features and thus making the interpretation of principal component difficult.

Because of these constraints, PCA may degrade the accuracy of a model. Some of the ways that PCA can degrade accuracy include [13]:

1. Loss of information: PCA is a linear technique that works by finding the directions in the data that explain the most variance. By reducing the dimensionality of the data, PCA can also remove important information that is not captured by the principal components.
2. Overfitting: When a PCA is used as a preprocessing step before training a model, it can introduce overfitting if the model is not regularized properly.
3. Non-linear relationships: since PCA is a linear technique, it may not be able to capture non-linear relationships between the input features. If the data has complex non-linear relationships, PCA may not be able to extract the most informative features, leading to a decrease in accuracy.
4. Outliers: PCA is sensitive to outliers, which can have a disproportionate effect on the principal components. If the data has outliers, PCA may remove important information from the data, leading to a decrease in accuracy [15].
5. It is also important to note that since PCA is an unsupervised technique, it does not consider the class information. Therefore, PCA-based dimensionality reduction may be suboptimal for supervised learning problems, where the goal is to find features that are informative for a specific task [16, 17].

In this research, we propose a dimension aggregation techniques to fix such problem such that accuracy increases.

2. Materials and Methods

Although the nature of the proposed methodology is generic, but here, it is presented for diagnosing breast cancer. For this, it considers a number of medical features. It comprises multiple information processing layers from feeding feature vector to preprocessing layer followed by PCA transformation, feature selection (dimensionality reduction), and features power set generation based data sampling and exploitation of each power set by an appropriate classification technique. Among these information-processing steps, the major novelty of the proposed technique lies in the power set generation of the feature dimensions. The flow diagram of the proposed system is shown in Figure 2. The following subsections describe these information processing layers.

2.1 Feature Vector

The first layer of the proposed technique is a vector representation of tumor based upon its statistical characteristics consisting eleven averages (radius, texture, perimeter, area, compactness, concavity, symmetry, fractal dimensions, smoothness, concave points, & symmetry error), five standard deviations (fractal dimension, radius, texture, perimeter, & area), and nine worst measurements (texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, & fractal dimension) as shown in Eq. (1) as follows:

$$\mathbf{x} = [x_1, x_2, \dots, x_n] \tag{1}$$

2.2 Data Normalization

The feature vectors, which comes from the dataset, have un-normalized data. Thereby, in preprocessing, normalization is performed. Unluckily, the features have different ranges, which may suffer classification. Thereby, after data encoding, normalization is performed on each feature to scale data with zero mean and unit variance according to Eq. **Error! Reference source not found.**) as given below:

$$x_o = \frac{x_i - \mu}{\sigma} \tag{2}$$

where, μ is mean and σ is standard deviation of the feature values. It is to note that this normalization is performed on each feature. Additionally, the same means and standard deviations computed in the training process to be used when normalizing a testing sample.

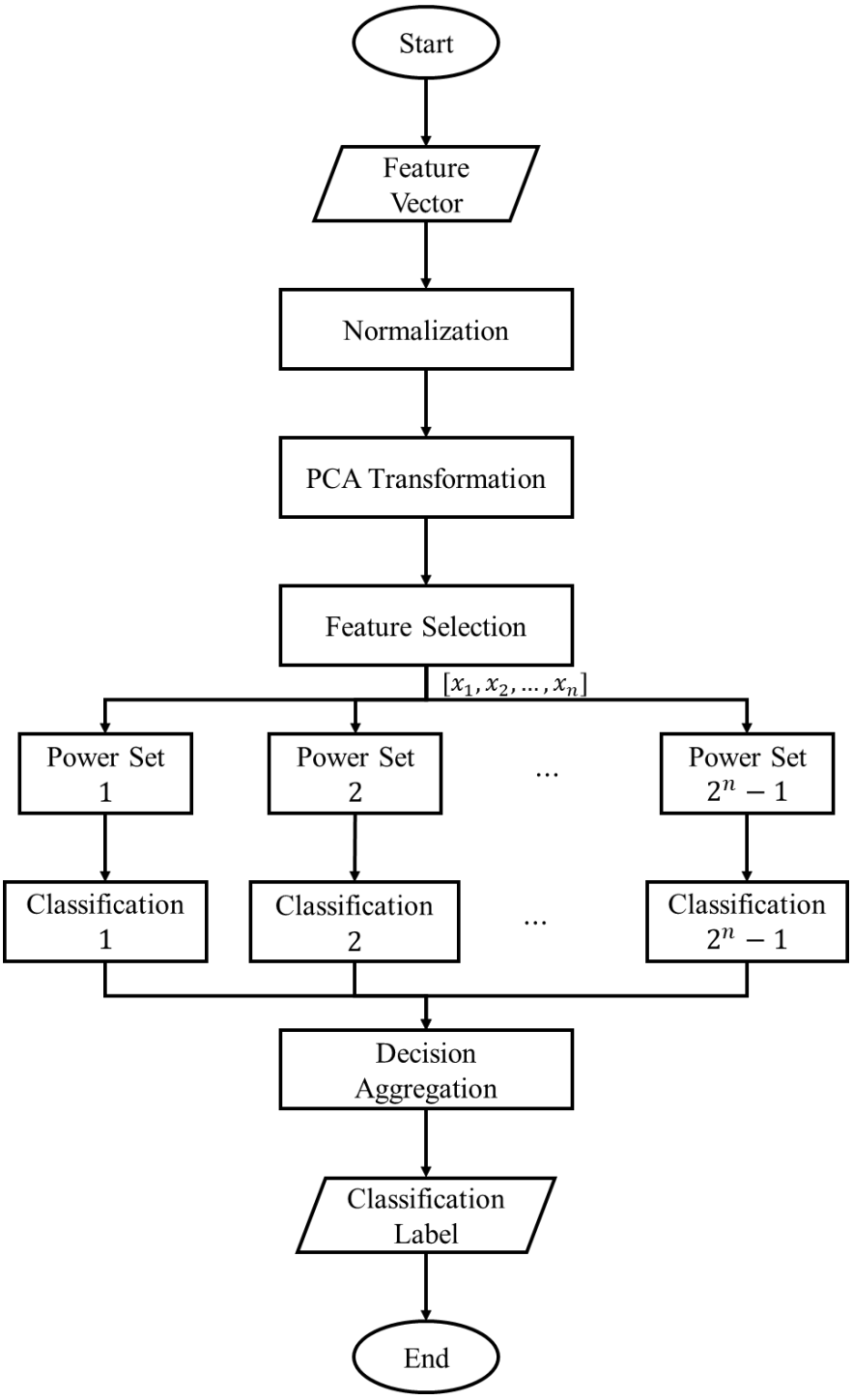


Figure 2. Flow diagram of the proposed methodology

2.3 Principal Component Transformation

Principal component analysis is although a famous technique used for feature reduction, but it may be ineffective if the dataset is uncorrelated. To check whether principal component analysis is suitable for feature reduction, correlation coefficient matrix of the dataset was computed according Eq. (3) as shown graphically in Figure 1. Since, the dataset contains good correlations, the employment of principal component analysis is suitable for it.

In the PCA decomposition of the dataset, it is to note that in the training of the proposed system, the whole dataset to be used for computation of principal components followed by the elimination of low variance principal components such that the chosen principal components covering at least 95% variance

of the whole dataset. Whereas, in testing of an unknown sample, already learnt principal components to be used for feature reduction.

$$[y_1, y_2, \dots, y_m] = \Im([x_1, x_2, \dots, x_n]) \tag{3}$$

$$\mathbf{P_x} = \begin{bmatrix} \frac{\mathrm{Cov}(x_1, x_1)}{\sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_1)}} & \frac{\mathrm{Cov}(x_1, x_2)}{\sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_2)}} & \cdots & \frac{\mathrm{Cov}(x_1, x_n)}{\sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_n)}} \\ \frac{\mathrm{Cov}(x_2, x_1)}{\sqrt{\mathrm{Var}(x_2)\mathrm{Var}(x_1)}} & \frac{\mathrm{Cov}(x_2, x_2)}{\sqrt{\mathrm{Var}(x_2)\mathrm{Var}(x_2)}} & \cdots & \frac{\mathrm{Cov}(x_2, x_n)}{\sqrt{\mathrm{Var}(x_2)\mathrm{Var}(x_n)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{Cov}(x_n, x_1)}{\sqrt{\mathrm{Var}(x_n)\mathrm{Var}(x_1)}} & \frac{\mathrm{Cov}(x_n, x_2)}{\sqrt{\mathrm{Var}(x_n)\mathrm{Var}(x_2)}} & \cdots & \frac{\mathrm{Cov}(x_n, x_n)}{\sqrt{\mathrm{Var}(x_n)\mathrm{Var}(x_n)}} \end{bmatrix} \tag{4}$$

2.4 Feature Selection

Based upon the usefulness of the principal components computed using Eq. (4), high variance components are to be selected for further processing. It is to note that the incorporation of principal components is only beneficial if there is a strong correlation between data dimensions. Thereby, if the data is already uncorrelated the employment of layer 2.3 and 2.4 is likely not to be useful. Thereby, in such cases, the proposed model is to be start continue to features' power set generation after normalization.

2.5 Features' Power Set Generation

After PCA based dimension reduction, the features are although compact and uncorrelated resulting in efficient training and testing but the classification performance may be decreased. This purpose of this layer is to exploit each possible combination of features for classification purpose. For this, the concept of the number of arrangements using combinations is employed. Thereby, selecting k features out of m features has $\binom{m}{k}$ possibilities. We choose $k \in 1,1,3, \dots, m$ and thus generating total $2^m - 1$ sub feature spaces as follows in Eq. (5):

$$\sum_{k=0}^{m} \binom{m}{k} = 2^m \tag{5}$$

From these $2^m - 1$ sub feature spaces, the probability of linear classification using a particular subspace having l dimensions can be computed using Cover's theorem [18] as follows in Eq. (6)

$$p_l = \frac{1}{2^{N-1}} \sum_{k=0}^{l} \binom{N-1}{k} \tag{6}$$

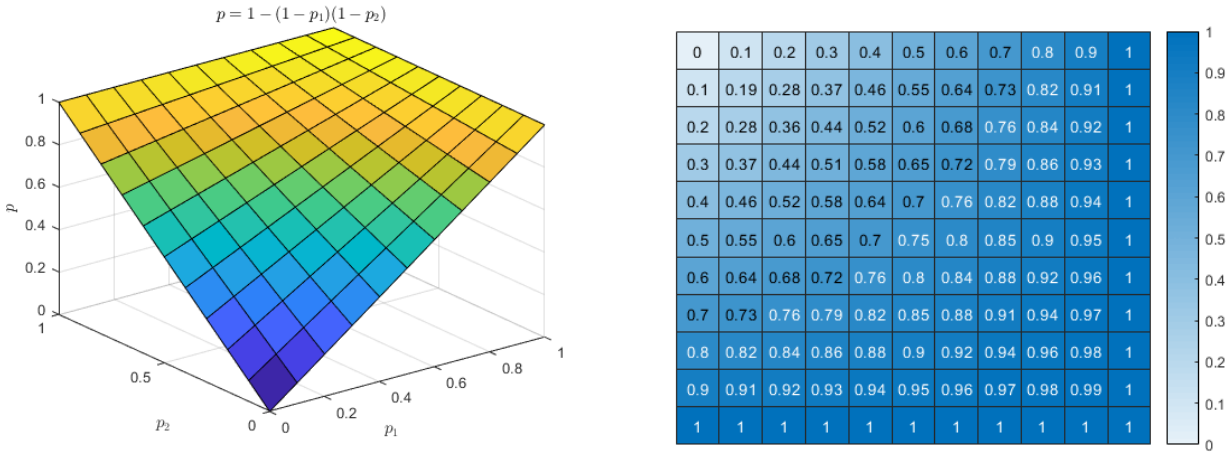
Using these probabilities, the probability of nonlinear classification is possible only if each l dimension subspace cannot provide linear classification i.e., the product of $\left(1 - P(l)\right)$ and thus, the probability of linear classification can be computed as follows in Eq. (7). The formulation shown in this equation has characteristics of being

$$p = 1 - \prod_{l=1}^{m} (1 - p_l) \tag{7}$$

Because of its computation from logical disjunction, it yields $p > p_l$ for all l. To understand it, consider $m = 2$ generating the probability as in Eq. (8)

$$p = 1 - (1 - p_1)(1 - p_2) = p_1 + p_2 - p_1 p_2 \tag{8}$$

Since, $0 \leq p_1, p_2 \leq 1$ thereby, $p_1, p_2 \leq p_1, p_2$ which implies that $p_1 - p_1 p_2 \geq 0$ and $p_2 - p_1 p_2 \geq 0$ and thus, Eq. (7) implies that $p \geq p_1, p_2$. It is to note that if $0 < p_1, p_2 < 1$ then $p > p_1, p_2$ i.e., p is always greater than p_1 and p_2. This is also illustrated in Figure 3 and Table 1.



0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.19	0.28	0.37	0.46	0.55	0.64	0.73	0.82	0.91	1
0.2	0.28	0.36	0.44	0.52	0.6	0.68	0.76	0.84	0.92	1
0.3	0.37	0.44	0.51	0.58	0.65	0.72	0.79	0.86	0.93	1
0.4	0.46	0.52	0.58	0.64	0.7	0.76	0.82	0.88	0.94	1
0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
0.6	0.64	0.68	0.72	0.76	0.8	0.84	0.88	0.92	0.96	1
0.7	0.73	0.76	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1
0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98	1
0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1
1	1	1	1	1	1	1	1	1	1	1

(a) (b)

Figure 3. Graphical illustration of the probability computed from Eq. (7) (a) plot and (b) tabular

Table 1. Comparison for various instances of individual and combined probability

Sr. #	p_1	p_2	$p = 1 - (1 - p_1)(1 - p_2)$
1	0.0	0.0	0
2	0.1	0.1	0.19
3	0.2	0.2	0.36
4	0.3	0.3	0.51
5	0.4	0.4	0.64
6	0.5	0.5	0.75
7	0.6	0.6	0.84
8	0.7	0.7	0.94
9	0.8	0.8	0.96
10	0.9	0.9	0.99
11	1	1	1

2.6 Dedicated Classification Layer

After the generation of $2^m - 1$ overlapping subspaces of the feature space, each subspace to be exploited by a dedicated base learner. For this, we employed bagging based combined classification based open majority voting of base learners generated by tree inducers. Such that data manipulation is provided by the power set generation of PCA dimensions as mentioned in Figure 2.

2.7 Decision Aggregation

The employment of dedicated classification for each data subset results in several labels for a testing sample. These labels need to be aggregated to transform several labels into one and final label. This component is aimed for this purpose. Thereby, when a test sample is processed by this component, the proposed model is logically finished.

3. Results

This section provides a detailed description of the experiments performed, evaluation metrics, and corresponding results. In addition, it also provides a brief description of the datasets used for the evaluation. Different ensemble models were used to find out which ensemble model performed better on our data.

3.1 Dataset

In this research work, a dataset regarding breast cancer was utilized. Wisconsin Diagnostic Breast Cancer (WDBC) Dataset is publicly available from the UCI machine learning repository. There are precisely 32 attributes in WDBC dataset and 569 instances. In these 32 features, one attribute is patient ID. All other attributes are deducted from a digitalized image that defines ten real-value features of the FNA sample calculated for each cell nucleus to determine whether a patient is diagnosed with malignant or benign. Three hundred fifty-seven cases were identified as "benign," and the remaining were classified as "malignant."

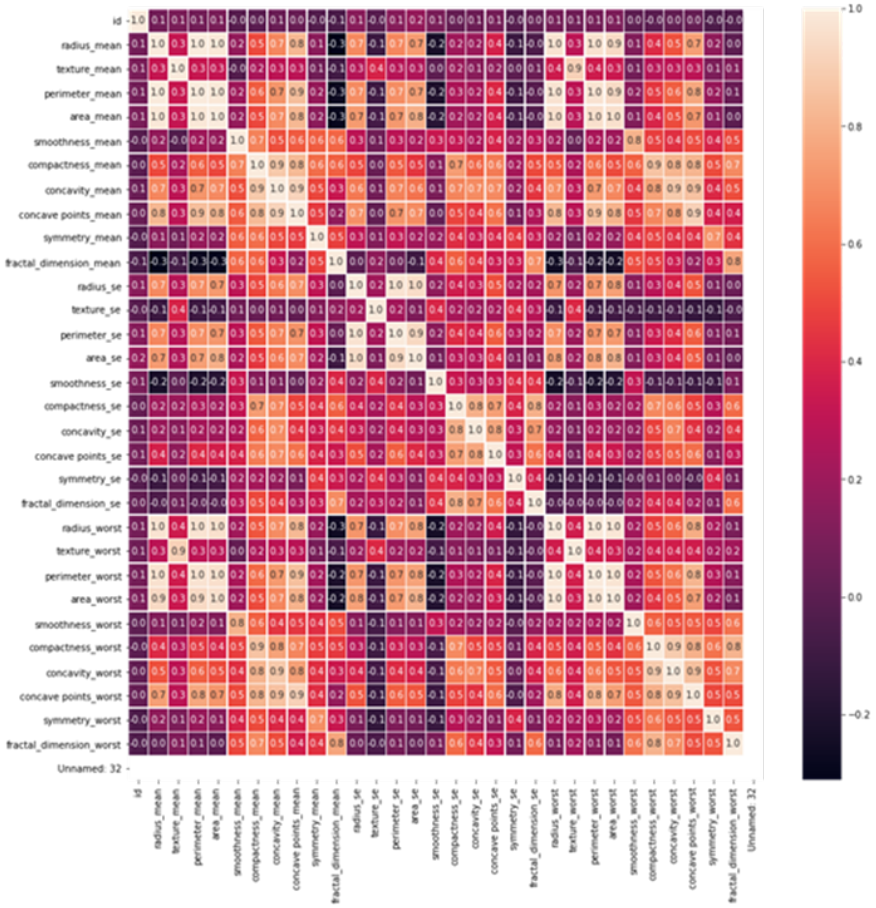


Figure 4. Graphical visualization of correlation coefficient matrix

3.2 Dataset Normalization

For normalization as stated in Eq. **Error! Reference source not found.**), mean and standard deviation of each feature were computed as shown in Table 2. It is to note that the same parameters computed in training phase will be used to normalize when processing an unknown sample to predict its label.

3.3 Principal Component Analysis

After preprocessing the dataset, we computed dataset correlation matrix as defined in Eq. (4) as visualized in Figure 4. By observing this matrix, it is to note that there are strong correlations among dataset dimensions thereby, it is beneficial to employ the principal component analysis to transform into uncorrelated dataset for further effective processing.

Table 2. Mean and standard deviation of each feature in the WDBC dataset

Sr. #	Feature	Mean (μ)	Deviation (σ)
01	mean radius	14.13	3.52
02	mean texture	19.29	4.30
03	mean perimeter	91.97	24.28
04	mean area	654.89	351.60
05	mean smoothness	0.10	0.01

06	mean compactness	0.10	0.05
07	mean concavity	0.09	0.08
08	mean concave points	0.05	0.04
09	mean symmetry	0.18	0.03
10	mean fractal dimension	0.06	0.01
11	radius error	0.41	0.28
12	texture error	1.22	0.55
13	perimeter error	2.87	2.02
14	area error	40.34	45.45
15	smoothness error	0.01	0.00
16	compactness error	0.03	0.02
17	concavity error	0.03	0.03
18	concave points error	0.01	0.01
19	symmetry error	0.02	0.01
20	fractal dimension error	0.00	0.00
21	worst radius	16.27	4.83
22	worst texture	25.68	6.14
23	worst perimeter	107.26	33.57
24	worst area	880.58	568.86
25	worst smoothness	0.13	0.02
26	worst compactness	0.25	0.16
27	worst concavity	0.27	0.21
28	worst concave points	0.11	0.07
29	worst symmetry	0.29	0.06
30	worst fractal dimension	0.08	0.02

Table 3. Illustration of generating the power set consisting subspaces of three ($m = 5$) PCA dimensional feature space

Subset #	**Principal Component's Presence**				
	y_1	y_2	y_3	y_4	y_5
0	0	0	0	0	0
1	0	0	0	0	1
2	0	0	0	1	0
3	0	0	0	1	1
4	0	0	1	0	0
5	0	0	1	0	1
6	0	0	1	1	0
7	0	0	1	1	1
8	0	1	0	0	0
9	0	1	0	0	1
10	0	1	0	1	0

11	0	1	0	1	1
12	0	1	1	0	0
13	0	1	1	0	1
14	0	1	1	1	0
15	0	1	1	1	1
16	1	0	0	0	0
17	1	0	0	0	1
18	1	0	0	1	0
19	1	0	0	1	1
20	1	0	1	0	0
21	1	0	1	0	1
22	1	0	1	1	0
23	1	0	1	1	1
24	1	1	0	0	0
25	1	1	0	0	1
26	1	1	0	1	0
27	1	1	0	1	1
28	1	1	1	0	0
29	1	1	1	0	1
30	1	1	1	1	0
$\mathbf{2^5 = 31}$	1	1	1	1	1

3.4 Feature Selection

After transforming the correlated dataset into uncorrelated dataset, we noted that first five principal components covering 85% of total variance, as shown in the Figure 5. It is to note that the number of input dimensions exponentially affects the number of power sets. If the number is large, it will enhance the complexity of the proposed model. This is the reason of choosing the minimum number of principal components, although they are not covered even 95% variance. However, the proposed model is so powerful that it is capable of boosting accuracy even with these few principal components as presented in the following subsections.

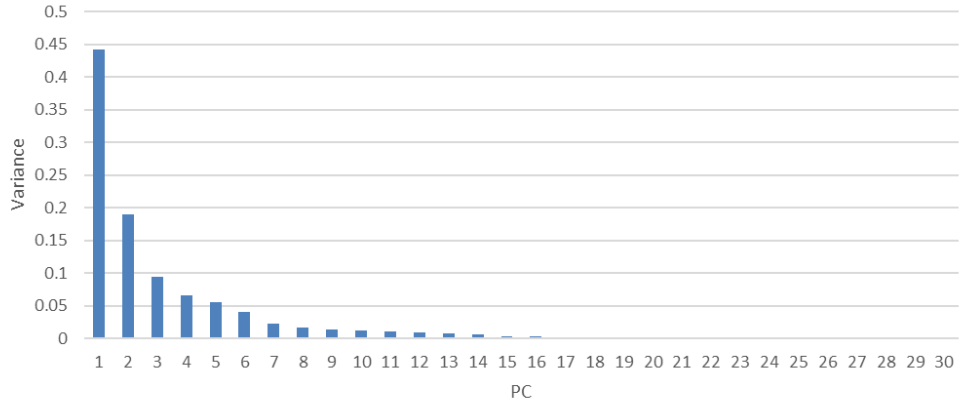


Figure 5. Variance of principal components versus prinfcipal components

3.5 Power Set Generation

Since we have chosen five principal dimensions, thereby, according to Eq. (5) the size of power set is $2^5 = 31$. The data subsets are achieved using the concept of binary counting as described in Table 3. For

example, the first data subset (0) is the empty set and it contains none of the input PCA dimension. The next subset (1) contains only fifth PCA dimension and so on.

3.6 Dedicated Classification

After generating power set consisting all (2^5) sub feature spaces, the next layer considers multiple base inducers such as random forest, Naïve Bayes, decision tree, and logistic regression. Because of the nature of dedicated classification for each subspace, this step resulted in $\mathbf{2^5}$ dedicated base learners for each inducer.

3.7 Decision Aggregation

The dedicated classification layer yields a number of output labels, one from each dedicated base learner. These decisions are combined using simple majority voting. However, if the problem is cost-sensitive, then majority voting may be replaced by minority voting of the classification labels. The classification confusion matrices of different classifiers are presented in Table 4. Whereas, the classification parameters are presented in Table 5 and ROC is shown in Figure 6.

Table 4. Confusion matrices of different classifiers before and after applying the proposed data sampling technique

Sr. #	Classifier	TP	FN	FP	TN
01	Decision Tree	62	2	8	100
02	Ensemble Decision Tree	46	1	1	66
03	Logistic Regression	45	2	2	65
04	Ensemble Logistic Regression	44	3	2	65
05	Naïve Bayes	42	5	6	61
06	Ensemble Naïve Bayes	42	5	3	64
07	Random Forest	61	2	7	101
08	Ensemble Random Forest	46	1	0	67

Table 5. Performance parameters of different classification techniques before and after applying the proposed data sampling technique

Sr. #	Classifier	Accuracy	Precision	Recall/TPR	TNR	FPR	FNR
01	Decision Tree	0.942	0.886	0.969	0.926	0.074	0.031
02	Ensemble Decision Tree	0.982	0.979	0.979	0.985	0.015	0.021
03	Logistic Regression	0.965	0.957	0.957	0.970	0.030	0.043
04	Ensemble Logistic Regression	0.956	0.957	0.936	0.970	0.030	0.064
05	Naïve Bayes	0.904	0.875	0.894	0.910	0.090	0.106
06	Ensemble Naïve Bayes	0.930	0.933	0.894	0.955	0.045	0.106
07	Random Forest	0.947	0.897	0.968	0.935	0.065	0.032
08	Ensemble Random Forest	0.991	1.000	0.979	1.000	0.000	0.021

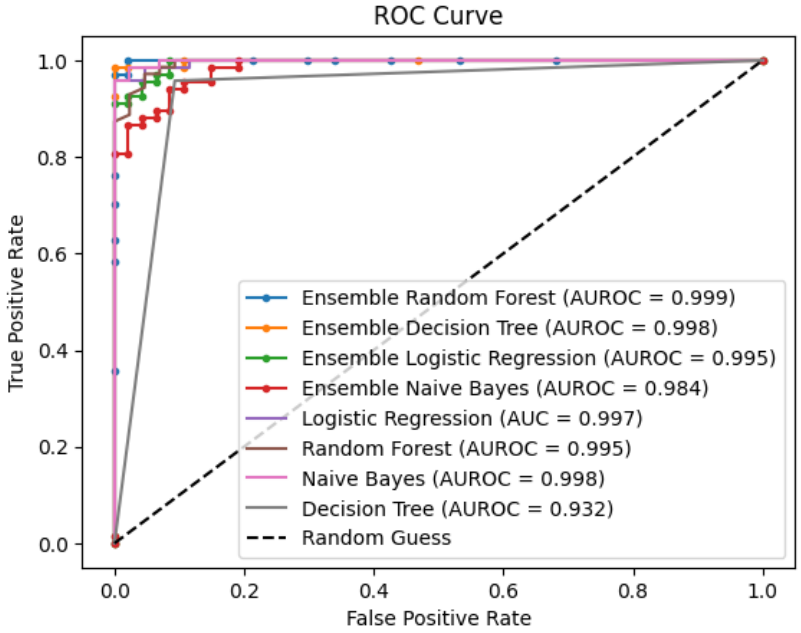


Figure 6. ROC curve based comparison between different classification techniques

4. Discussion

An effective classification is represented by its capability to differentiate various classes more accurately and more precisely. Other than accuracy and precision, various other parameters such as true positive rate, true negative rate, false positive rate, and false negative rate are also used to assess the effectiveness of classification. It can be noticed by observing all performance measure presented in Table 5 that the proposed data sampling technique inspired by the mechanism of generating subsets of power set resulted in more accurate and more precise classification. This has been possible because of the nature of data sampling which extracted a number of feature spaces (perspectives|) empowered by the dedicated classifications. Combing the individual decisions of a number of dedicated classifiers for ensemble classification makes the overall classification capable of exploiting effected a number of feature spaces and thus boosting the classification performance parameters. For example, consider row 7 and 8 of Table 5 representing the classification parameters of random forest before and after the proposed data sampling based ensemble classification. It is observed that the proposed data sampling based ensemble classification resulted in boosting accuracy from 0.947 to 0.991, precision from 0.897 to 1, true positive rate from 0.968 to 0.979, and true positive rate from 0.935 to 1. On the other hand, it resulted in a reduction of false positive rate from 0.065 to 0 and false negative rate from 0.032 to 0.021.

5. Conclusions

This research successfully considered the challenging problem of classification for a large dimensional feature space. The reason of the difficulty in such feature spaces is because; exploiting a large dimensional feature space is associated with deducing a decision boundary from the dataset. Since, decision boundary is described by the parameters whose count is proportional to the number of feature dimensions and thus, classification in large dimensional feature space requires a large number of parameters to be optimized. Unluckily, if the number of instances in the dataset is relatively small as compared to its feature dimension then this optimization (training process) is challenging which result in classification with constrained accuracy. Additionally, this process also resulted in time taking training process. This research considers this problem and it is motivated by the generation of power set of set theory. For this, it considers the dimensions of feature space as elements of set and generates all subsets to build the power set. Although this research employs principal component analysis, but it is only recommended subject to the presence of correlated feature space. The proposed technique is highly beneficial for the large dimensional feature spaces since this particular feature sampling increases the probability of effective classification according

to Cover's theorem as validated by the experiments. However, for low dimensional feature space the proposed technique may behave inefficacy.

Data Availability Statement: The authors declared that the datasets used in this research are publicly available.
Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

1. H. Abdel-Razeq et al., "Differences in clinicopathological characteristics, treatment, and survival outcomes between older and younger breast cancer patients," Scientific Reports, vol. 11, no. 1, p. 14340, 2021/07/12 2021, doi: 10.1038/s41598-021-93676-w.
2. J. F. M. Hyuna Sung PhD, ME, Rebecca L. Siegel MPH, Mathieu Laversanne MSc, Isabelle Soerjomataram MD, MSc, PhD, Ahmedin Jemal DMV, PhD, Freddie Bray BSc, MSc, PhD, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209-249, 2021.
3. M. B. Saima Shakil Malik, Muhammad Bilal Khan, Nosheen Masood, "Survival analysis of breast cancer patients with different treatments: a multicentric clinicopathological study," Journal of Pakistan Medical Association, vol. 69, no. 7, pp. 976-980, 2019.
4. A. G. Waks and E. P. Winer, "Breast Cancer Treatment: A Review," JAMA, vol. 321, no. 3, pp. 288-300, 2019, doi: 10.1001/jama.2018.19323.
5. X. Yu, Q. Zhou, S. Wang, and Y.-D. Zhang, "A systematic survey of deep learning in breast cancer," International Journal of Intelligent Systems, vol. 37, no. 1, pp. 152-216, 2022, doi: https://doi.org/10.1002/int.22622.
6. N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," Computer Methods and Programs in Biomedicine, vol. 156, pp. 25-45, 2018/03/01/ 2018, doi: https://doi.org/10.1016/j.cmpb.2017.12.012.
7. GLOBOCAN 2020, International Agency for Research on Cancer 2021, World Health Organization, United Nations.
8. M. Björklund, "Be careful with your principal components," Evolution, vol. 73, no. 10, pp. 2151-2158, 2019, doi: https://doi.org/10.1111/evo.13835.
9. E. Elhaik, "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," Scientific Reports, vol. 12, no. 1, p. 14683, 2022/08/29 2022, doi: 10.1038/s41598-022-14395-4.
10. J. Stuckman, J. Walden, and R. Scandariato, "The Effect of Dimensionality Reduction on Software Vulnerability Prediction Models," IEEE Transactions on Reliability, vol. 66, no. 1, pp. 17-37, 2017, doi: 10.1109/TR.2016.2630503.
11. C. Gambella, B. Ghaddar, and J. Naoum-Sawaya, "Optimization problems for machine learning: A survey," European Journal of Operational Research, vol. 290, no. 3, pp. 807-828, 2021/05/01/ 2021, doi: https://doi.org/10.1016/j.ejor.2020.08.045.
12. S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A Survey of Optimization Methods From a Machine Learning Perspective," IEEE Transactions on Cybernetics, vol. 50, no. 8, pp. 3668-3681, 2020, doi: 10.1109/TCYB.2019.2950779.
13. E. Elhaik, "Why most Principal Component Analyses (PCA) in population genetic studies are wrong," bioRxiv, p. 2021.04.11.439381, 2021, doi: 10.1101/2021.04.11.439381.
14. M. Scholz, "Validation of Nonlinear PCA," Neural Processing Letters, vol. 36, no. 1, pp. 21-30, 2012/08/01 2012, doi: 10.1007/s11063-012-9220-6.
15. E. Gañan-Cardenas and J. C. Correa-Morales, "Comparison of Correction Factors and Sample Size Required to Test the Equality of the Smallest Eigenvalues in Principal Component Analysis," Revista Colombiana de Estadística, vol. 44, pp. 43-64, 2021. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-17512021000100043&nrm=iso.
16. T. Howley, M. G. Madden, M.-L. O'Connell, and A. G. Ryder, "The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data," in Applications and Innovations in Intelligent Systems XIII, London, A. Macintosh, R. Ellis, and T. Allen, Eds., 2006// 2006: Springer London, pp. 209-222.
17. A. Cheriyadat and L. M. Bruce, "Why principal component analysis is not an appropriate feature extraction method for hyperspectral data," in IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477), 21-25 July 2003 2003, vol. 6, pp. 3420-3422 vol.6, doi: 10.1109/IGARSS.2003.1294808.
18. T. M. Cover, "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," IEEE Transactions on Electronic Computers, vol. EC-14, no. 3, pp. 326-334, 1965, doi: 10.1109/PGEC.1965.264137.