Comparative analysis of various non-linear dimensionality reduction techniques for cervical cancer prediction.

Anjor Rao[1] and Rabiya Waseem[2]

[1]Applied CS, University of Winnipeg, 515 Portage Ave, Winnipeg, R3B 2E9, Manitoba, Canada.
[2]Applied CS, University of Winnipeg, 515 Portage Ave, Winnipeg, R3B 2E9, Manitoba, Canada.

Contributing authors: rao-a@webmail.uwinnipeg.ca; waseem-r@webmail.uwinnipeg.ca;

Abstract

Cervical cancer is a foremost cancer cause of death in women around the world. Each year in Canada,1300 plus women are diagnosed with cancer and over 400 die from the disease. Early detection and prevention can help to reduce mortality. There are numerous factors such as smoking, age of patient, number of pregnancies can help to detect whether a person has cancer or not. Anticipating the cancer diagnosis using a machine learning model, can provide great accuracy and mitigate the mortality rate. In our proposal, we will be implementing non-linear dimensionality reduction techniques such as Multidimensional Scaling (MDS), Isometric mapping (ISOMAP), Locally-linear embedding (LLE) ,T-Distributed stochastic neighbor embedding (T-SNE) on classification models i.e. Logistic regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF), and evaluate and analyze the results generated by each model.

Keywords: Cervical cancer ,non-linear dimensionality reduction, t-sne, lle, mds, isomap, classification, SMOTE

1 Introduction

Cervical cancer causes around 260,000 deaths per annum and around 85% of these deaths are found in developing countries. It is one of the major causes of cancer deaths in women. Imbalance of proper health and deprivation causes a major role in this high mortality rate. The World Health Organization (WHO) estimates that of the 500,000 new cases annually, 80% affect women between the ages of 15 and 45 years who live in developing nations. It predominantly impacts women living in Latin America and the Caribbean, sub-Saharan Africa, and Southeast Asia. Only 5% of women in these regions have been screened for cervical disease in the past 5 years (Fig.1) [1]. Smoking is considered a crucial cause of cervical cancer. Multiple pregnancies is also one of major factors causing the cancer. Early stage symptoms are difficult to detect. The symptoms are visible after a few stages.

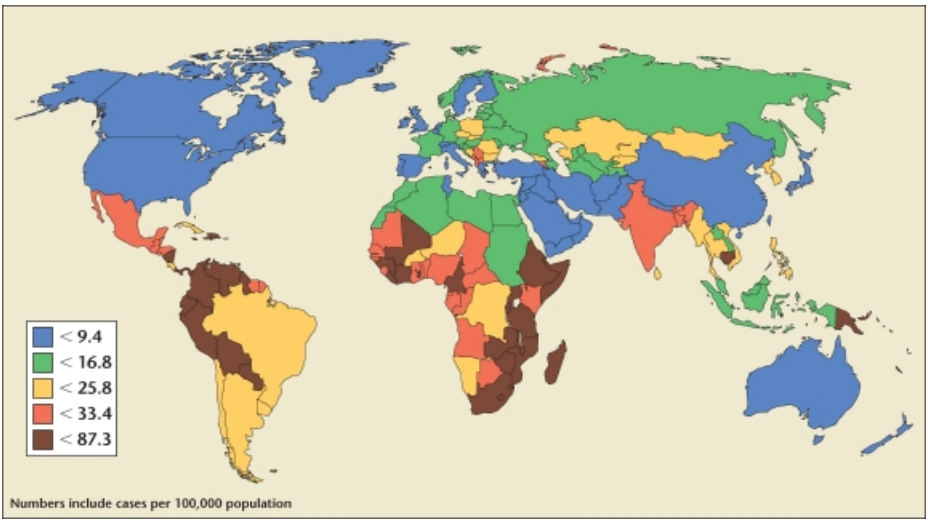

Fig. 1 Global burden of cervical cancer greatest in developing countries. Reproduced with permission from Women

Screening and different deterministic tests confuse the available Computed Aided Diagnosis (CAD) to treat the patient correctly for the cancer. The screening detected number of factors such as age, number of sexual partners, age of first sexual intercourse, number of pregnancies, smoking habits, hormonal, STD's in the patient. There are few screening methods like Schiller, Hinselmann and the standard biopsy test for cancer detection [2]

The machine learning techniques can help to anticipate whether a patient has cancer or not. In our proposal, we will provide an insight and analysis of results of different classification models like Logistic regression (LR), K-Nearest

Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF) using non-linear dimensionality reduction techniques such as Isometric mapping (ISOMAP), Locally-linear embedding (LLE), Multidimensional Scaling (MDS) and T-Distributed stochastic neighbor embedding (T-SNE).

2 Background

In [2] Parikh and Menon implemented three classifier models i.e K-nearest neighbor, decision tree classifier and random forest tree on UCI Cervical Cancer (Risk Factors) dataset for the biopsy test. The K-nearest neighbor scored the highest accuracy 95.3% while the accuracy of the decision tree classifier is 85.11%.

In 2020, [3] has proposed an analysis on the results based on various classification models such as Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest Decision Tree (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR). They primarily focused on the Biopsy test as the main target class. To handle the class imbalance problem, they implemented synthetic minority oversampling technique (SMOTE). The results displayed that the RF (96.40%) performed best as compared to kNN(92.60%) and SLR(76.90%).

[4] Choudhury performed similar analysis on different classification models with an oversampling method to handle the class imbalance. The k-Nearest Neighbors (KNN) (90.68%) performed well, while SVM and Logistic regression scored an accuracy of 81.67% and 63.97% respectively. However, non-linear dimensionality reduction has not been implemented before on this data set.

Dissanayake and Johar [5] did a comparative study on other dataset - heart disease, they implemented ten feature selection techniques i.e. ANOVA, Chi-square, mutual information, ReliefF, forward feature selection, backward feature selection, exhaustive feature selection, recursive feature elimination, Lasso regression, and Ridge regression, and six classification approaches, i.e., decision tree, random forest, support vector machine, K-nearest neighbor, logistic regression, and Gaussian Naive Bayes. The experimental finding suggested that the backward selection feature with the decision tree classifier provided highest accuracy of 88.52%.

In 2019, [6] ROOPA, ASHA implemented a regression model for predicting disease based on Principal Component Analysis (PCA). The accuracy obtained by this combination is 82.1%, which is greater than other basic regression models.

3 Theoretical Framework

3.1 Basic classification models

In our proposal, we have used four classification models i.e. Logistic regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF). Following are details of classification models

3.1.1 Logistic regression

LR is a statistical model that models the probability of one event occurring by having the log of odds for the event. In this model, there are one or more independent variables and single dependent variable. The binary logistic regression consists of 0 and 1 in the dependent variable column.

3.1.2 K-Nearest Neighbors

KNN is a supervised learning classification model. It classifies and assigns the class to new data points according to the similar class nearby within a given threshold. The threshold is the number of data points to be considered in the neighborhood. It is also called a lazy learner algorithm.

3.1.3 Support Vector Machine

SVM is a supervised machine learning algorithm. The model is used for both classification and regression problems. The classification is formed on the hyperplane that separates the two classes.

3.1.4 Random Forest

RF is a supervised machine learning algorithm used for classification and regression models. It builds multiple decision trees on a number of samples and takes their highest vote for the classification.

3.2 Introduction to non-linear dimensionality reduction

The machine learning dataset consists of thousands of features. Training a machine learning model with a large number of features can be a sluggish process, and also there are high chances of overfitting the data because there will be additional anticipation as compared to low-dimensional data. This problem can be called the "curse of dimensionality". To avoid the problem of huge numbers of dimensionalities, "Dimensionality reduction" techniques are used. It is a process of mapping high dimensional data to low-dimensional extracting crucial feature information. The dimensions can be reduced using two ways i.e. projection and manifold learning. According to mathematical terminology, a manifold is a topological space that simulates Euclidean space. The linear dimensionality reduction techniques like PCA are suitable for linear datasets. However, if the shape of the dataset is different like swiss roll etc. then the

PCA technique won't work in this case. Thus Manifold Learning solves the non-linear dataset problems in a systematic way. Following are the non-linear dimensionality reduction techniques we will be using in our approach:

- Multidimensional Scaling (MDS)
- T-Distributed stochastic neighbor embedding (T-SNE)
- Isometric mapping (ISOMAP)
- Locally-linear embedding (LLE)

3.3 Non-linear dimensionality reduction algorithms

3.3.1 Multidimensional Scaling

MDS is the analysis of proximity of a data set to identify the obscure structure of the data. This technique is further classified into two ways i.e. metric and non-metric. In both the cases, each object is characterized by either distances between or dissimilarities to others. Intuitively, the notion of distance, or dissimilarity, is very easy to understand: the dissimilarity is zero for identical objects and grows as they become increasingly different from each other. Conversely, similarity is high for nearly identical objects and decreases as differences appear. The equation of stress is formulates as below [7].

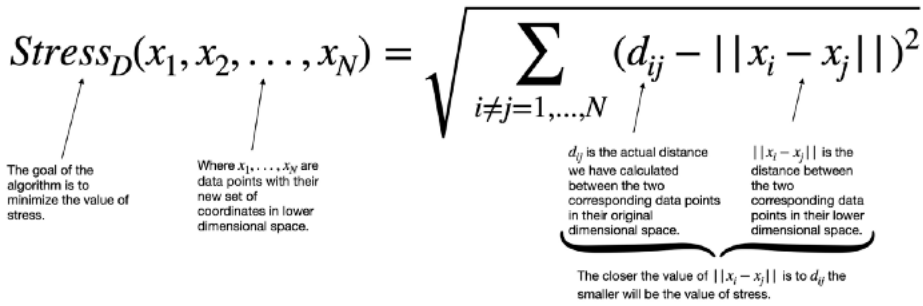

Fig. 2 MDS Equation 1: Stress value

Algorithm 1 Multidimensional Scaling algorithm

(I) Calculate the distances between each pair of points.

(II) With the original distances known, find a set of coordinates in a lower-dimensional space that minimizes the value of stress (MDS Equation 1).

3.3.2 T-Distributed stochastic neighbor embedding

T-SNE is a methodology for visualizing high-dimensional data into two or three dimensional maps by providing each data point a location. This technique is classified under non-linear dimensionality reduction technique. The similar data points are close together in lower dimensional space. The algorithm works using a local approach, i.e. mapping of nearby data points on the manifold to nearby data points in low dimensional space [8].

Algorithm 2 T-Distributed stochastic neighbor embedding algorithm

(I) Constructs a probability distribution over pairs of high-dimensional objects.

(II) t-SNE defines a similar probability distribution over the points in the low-dimensional map.

(III) It minimizes the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.

3.3.3 Isometric mapping

ISOMAP is the extended version of MDS. The only difference between the two methods is the metric used to measure the pairwise distances: Isomap uses graph distances instead of Euclidean ones in the algebraical procedure of metric MDS. Just by introducing the graph distance, the purely linear metric MDS becomes a nonlinear method [9].

Algorithm 3 Isometric mapping algorithm

(I) Construct the neighborhood graph G between the data point's X_i and X_j, if X_i is the K nearest neighbor of X_j or they are closer than particular distance ϵ where ϵ is the max Euclidean search distances E and K is the number of nearest neighbors.

(II) Calculate the geodesic distance matrix between all pairs of data points X_i and X_j using the shortest path distance graph G and then compute the shortest path distance by using Dijkstra's and Floyd's algorithms.

(III) Apply MDS to the resulting geodesic distance matrix to identify a low d-dimensional embedding by executing Eigendecomposition.

3.3.4 Locally-linear embedding

LLE is similar to ISOMAP. It works on local structure rather than preserving the global structure [10]

In research [10], minimizing functions for weights are formulated. The equations are as belows:

$$\mathcal{E}(W) = \sum_i \left| \vec{X}_i - \sum_j W_{ij} \vec{X}_j \right|^2,$$

Fig. 3 LLE Equation 1: Cost minimizing function

$$\Phi(Y) = \sum_i \left| \vec{Y}_i - \sum_j W_{ij} \vec{Y}_j \right|^2.$$

Fig. 4 LLE Equation 2: Cost minimizing function

Algorithm 4 Locally-linear embedding algorithm

(I) Compute the neighbors of each data point $\vec{X}_i$

(II) Compute the weights $\vec{W}_{ij}$ that best reconstruct each data point $\vec{X}_i$ from its neighbors, minimizing the cost using (LLE Equation 1) by constrained linear fits.

(III) Compute the vectors $\vec{Y}_i$ best reconstructed by the weights $\vec{W}_{ij}$, minimizing the quadratic form in (LLE Equation 2) by its bottom non zero eigenvectors

4 Implementation

In our proposal, we followed different stages i.e. data gathering, data preprocessing, choosing model, training and prediction to create a model.
The UCI Machine Learning Repository has a dataset named 'Cervical cancer (Risk Factors) Data Set' [11]. This dataset gives an emphasis on predicting whether a patient has cervical cancer or not based on multiple risk factors. There are 858 numbers of instances and 36 attributes. The attributes consist of 32 independent variables and 4 dependable variables i.e target variables. The target variables (Hinselmann, Schiller, Cytology, Biopsy) are test names for detection of cancer. In our proposal we are conducting classification for

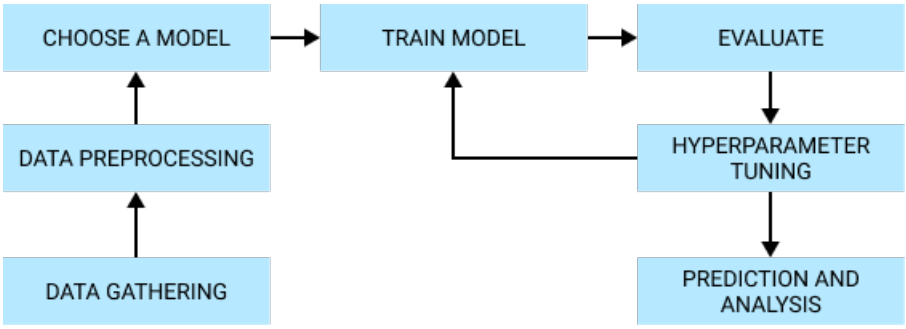

Fig. 5 Model flowchart

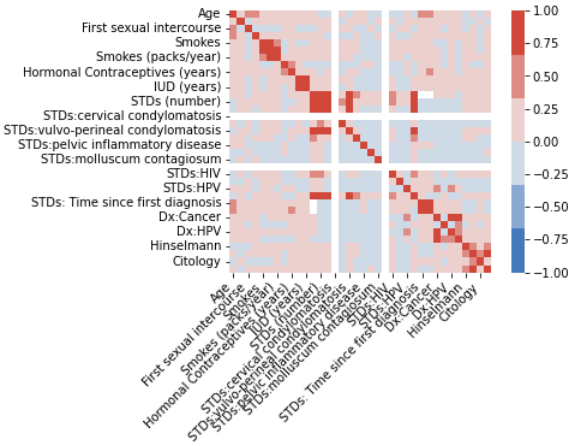

Fig. 6 Correlation heatmap between attributes

the Biopsy test since it is one of the most popular and recommended tests for cancer detection.

In the data preprocessing stage, initially we plotted correlation heatmap (Figure 6) to understand correlation between attributes. From the heatmap, it is evident that few features have high correlation like Age and First Sexual intecourse or Age and Smokes, while Age and STDs types are poor correlated. Further we looked at the data type and missing values and the description table is given below in (Table 1)

The feature attributes 'STDs: Time since first diagnosis' and 'STDs: Time since last diagnosis' contain up to 92% missing data. We discarded these columns from our dataset.

There are many integer columns containing missing values like 'First sexual

Attribute	Type	Missing Value Yes/No (%)
Age	Integer	No
Number of sexual partners	Integer	Yes (3%)
First sexual intercourse (age)	Integer	Yes (1%)
Number of pregnancies	Integer	Yes (7%)
Smokes	Boolean	Yes (2%)
Smokes (years)	Integer	Yes (2%)
Smokes (packs/year)	Integer	Yes (2%)
Hormonal Contraceptives	Boolean	Yes (13%)
Hormonal Contraceptives (years)	Integer	Yes (13%)
IUD	Boolean	Yes (14%)
IUD (years)	Integer	Yes (14%)
STDs	Boolean	Yes (12%)
STDs (number)	Integer	Yes (12%)
STDs:condylomatosis	Boolean	Yes (12%)
STDs: cervical condylomatosis	Boolean	Yes (12%)
STDs:vaginal condylomatosis	Boolean	Yes (12%)
STDs:vulvo-perineal condylomatosis	Boolean	Yes (12%)
STDs:syphilis	Boolean	Yes (12%)
STDs:pelvic inflammatory disease	Boolean	Yes (12%)
STDs:genital herpes	Boolean	Yes (12%)
STDs:molluscum contagiosum	Boolean	Yes (12%)
STDs:AIDS	Boolean	Yes (12%)
STDs:HIV	Boolean	Yes (12%)
STDs:Hepatitis B	Boolean	Yes (12%)
STDs:HPV	Boolean	Yes (12%)
STDs:Number of diagnosis	Integer	No
STDs:Time since first diagnosis	Integer	Yes (92%)
STDs:Time since last diagnosis	Integer	Yes (92%)
Dx:Cancer	Boolean	No
Dx:CIN	Boolean	No
Dx:HPV	Boolean	No
Dx	Boolean	No
Biopsy	Boolean	No

Table 1 Attributes, data type and missing value in data set

intercourse' and 'Hormonal Contraceptives (years)'. We visualized each column in boxplot and graph (Figure 8) (Figure 9)
From the visualization, we concluded that the data is skewed and has multiple outliers, thus we filled up the columns with median value.
Similarly for missing boolean values in columns like 'Smokes' or 'Hormonal Contraceptives' we replaced the missing values with mode value. (Figure 10)

The dataset has a class imbalance problem.The class imbalance problem is where the classes are skewed or biased to a particular class. There are multiple reasons for class imbalance like inaccurate collection of data. In our case, many patients choose to not answer a few questions due to privacy concerns [11]. We

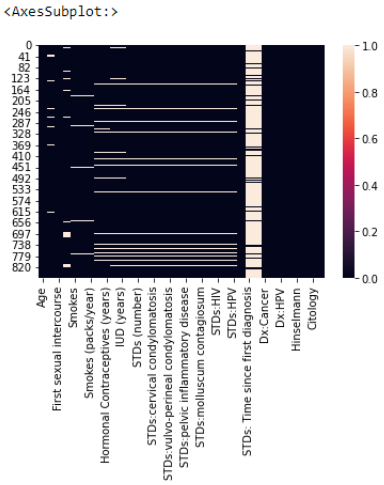

Fig. 7 Visualization of missing features in features

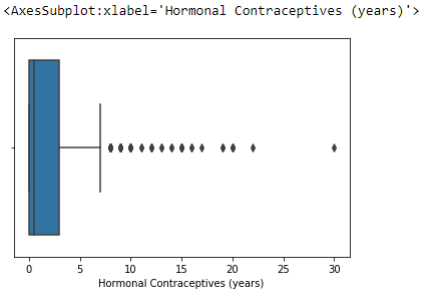

Fig. 8 Data Visualization of column 'Hormonal Contraceptives (years)' in box-plot

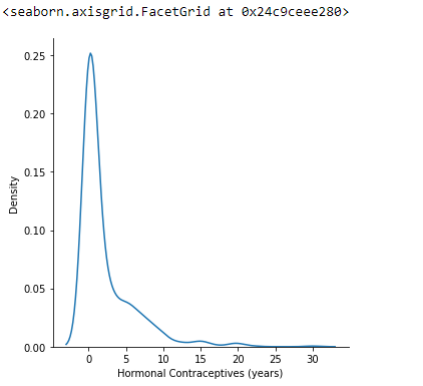

Fig. 9 Data Visualization of column 'Hormonal Contraceptives (years)' in graph

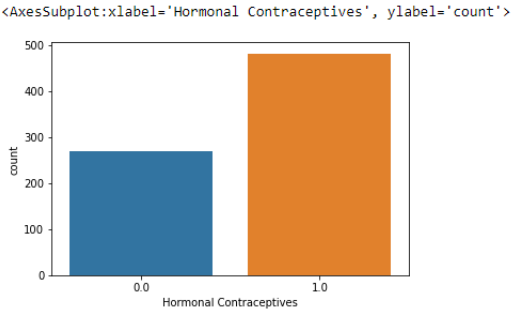

Fig. 10 Data Visualization of column 'Hormonal Contraceptives' in bar chart

handled this problem by using the SMOTE (Synthetic Minority Oversampling Technique). SMOTE is an oversampling method where synthetic samples are generated using minority classes.
Further static splitting of data sets for training and testing would not cover all the cases due to class imbalance problems. Thus we rectified this issue using K-Fold cross validation. In K Fold cross validation, the actual samples are partitioned into k equal size of sub samples, maintaining class imbalance through each sub sample. In our case we have considered 10 folds.
We will create a 'BASE' model without any dimensionality reduction technique applied on it, to compare the results. In the next step, we applied dimensionality reduction technique to the dataset and reduced 30 features into n-number of features i.e (components). Followed by training of classification models.
During the hyper parameter tuning stage, we increased component numbers of two to six for each dimensionality reduction method.

From (Figure 11) and (Figure 12) for ISOMAP, the accuracy and f1 score is good for two components, similarly for LLE, T-SNE and MDS accuracy and f1 score is good for six components.
Information about tools and libraries used:

- Python library: Scikit-learn library, Seaborn, Numpy, pandas, Matplotlib
- IDE: Jupyter Notebook
- Computing software: Anaconda Navigator 2.1.1
- Operating system: Windows 10

5 Experiments and results

The evaluation metrics for the models are accuracy, precision and recall which is depicted in the confusion matrix. These metrics are used to calculate the performance of our classification models on the risk factor of cervical cancer data between actual and predicted results. This analysis of actual output and predicted output is illustrated in confusion matrix shown in (Figure 13)

Where class 1 and 2 can be positive or negative. The terms can be described as follows:

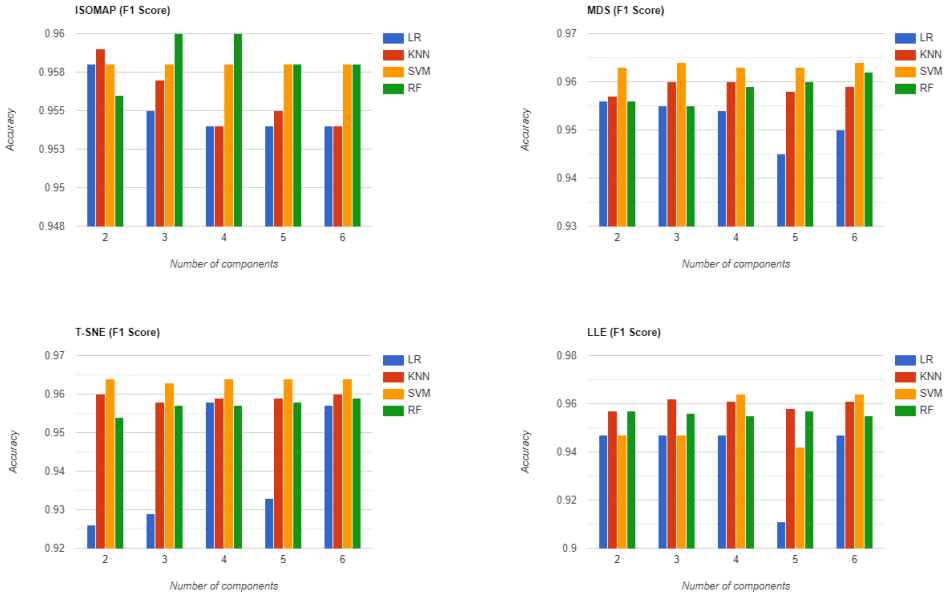

Fig. 11 Comparison of F1 score of ISOMAP, MDS, T-SNE, LLE in terms of components

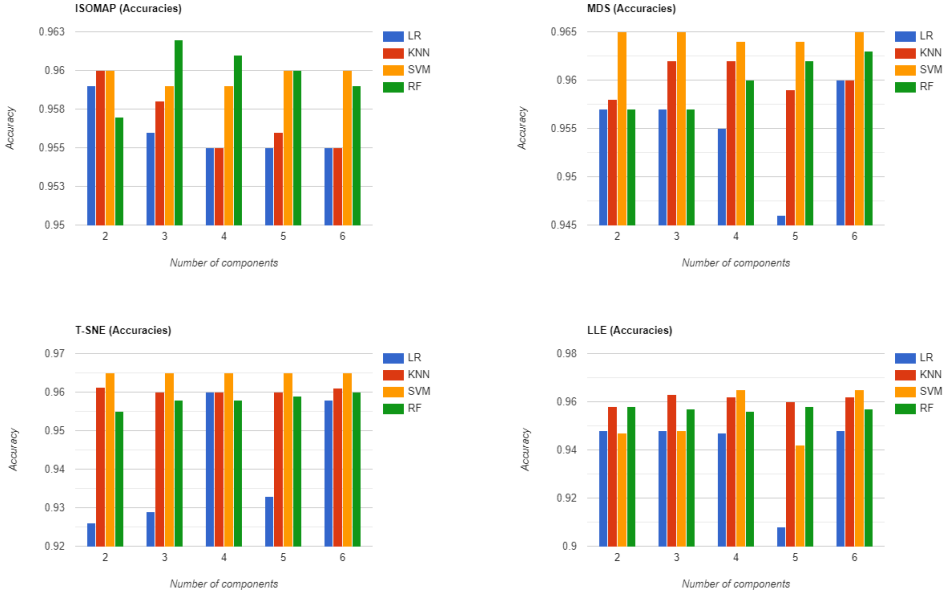

Fig. 12 Comparison of accuracies of ISOMAP, MDS, T-SNE, LLE in terms of components

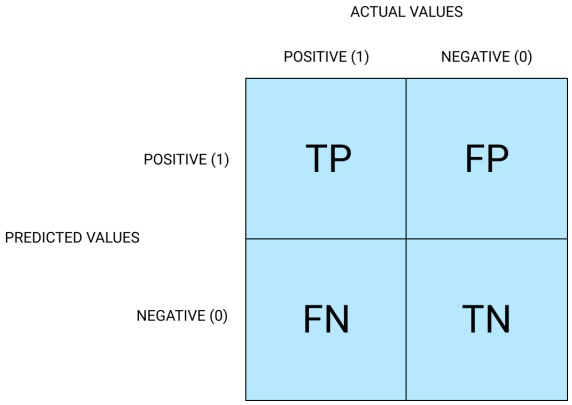

Fig. 13 Confusion matrix

Positive (P) : The observation is positive, for instance cancer detected positive.
Negative (N): The observation is negative, for instance cancer detected negative.
True Positive (TP): Observation is positive, and is predicted to be positive.
False Negative (FN): Observation is positive, but is predicted negative.
True Negative (TN): Observation is negative, and is predicted to be negative.
False Positive (FP): Observation is negative, but is predicted positive.

Accuracy can be defined as correct prediction if patient has cancer or not

Precision can be defined as where the total number of correctly classified positive samples are divided by the total number of actual positive samples.

Recall can be defined as where the ratio of the total number of correctly classified positive samples divided by the predicted total number of positive samples.

F1 score can be defined as harmonic mean that combines precision and recall Accuracy, Precision, Recall and F1 score can be formulated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

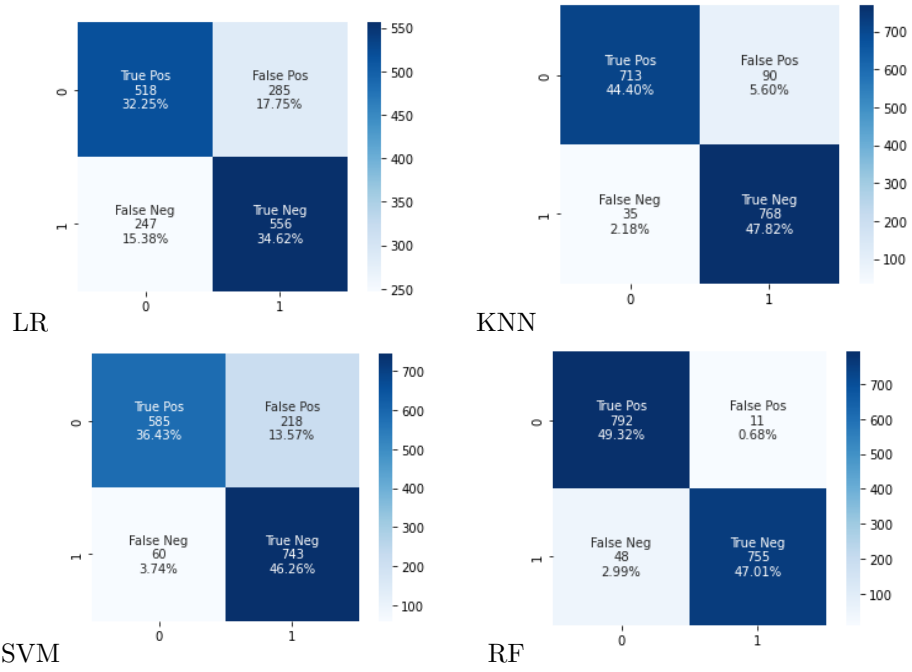

Fig. 14 Confusion matrices of BASE model

$$F1score = 2x\frac{precisionxrecall}{precision+recall}$$

5.1 Evaluation of Base model

5.1.1 Confusion matrices

The confusion matrices of BASE model is given in (Figure 14)

5.1.2 Metrics

The performance metrics of BASE model is given in (Table 2)

Metric	LR	KNN	SVM	RF
Accuracy	66.8%	92.2%	82.6%	96.3%
Precision	66.1%	89.5%	77.3%	98.5%
Recall	69.2%	95.6%	92.5%	94.0%
F1 score	67.6%	92.4%	84.2%	96.2%

Table 2 Performance metrics of BASE model

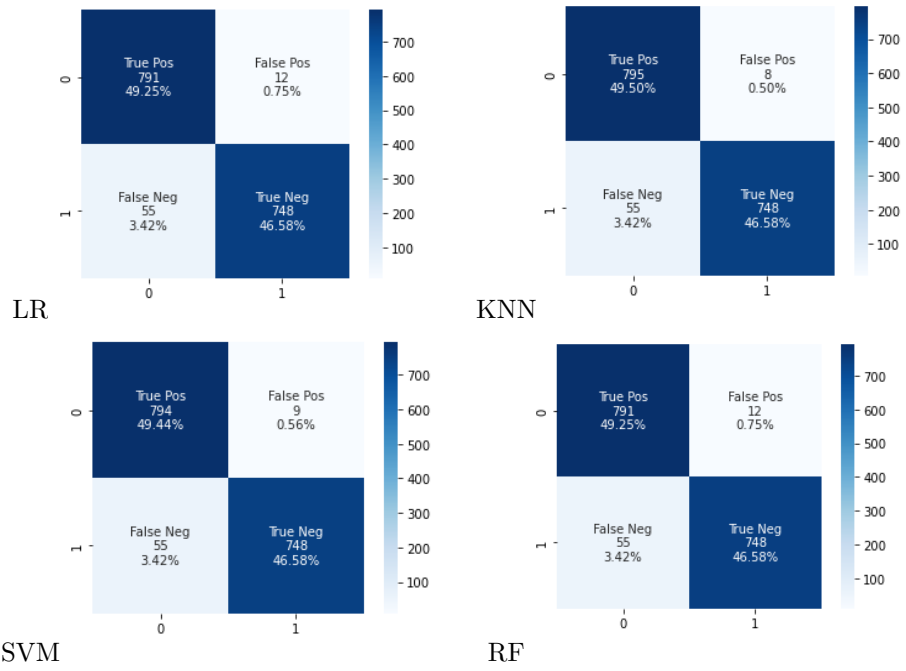

Fig. 15 Confusion matrices of models using ISOMAP method

5.2 Evaluation of ISOMAP

5.2.1 Confusion matrices

The confusion matrices of ISOMAP is given in (Figure 15)

5.2.2 Metrics

The performance metrics of model using ISOMAP is given in (Table 3)

Metric	LR	KNN	SVM	RF
Accuracy	95.8%	96.0%	96.0%	95.8%
Precision	98.4%	98.9%	98.8%	98.4%
Recall	93.1%	93.1%	93.1%	93.1%
F1 score	95.7%	95.9%	95.8%	95.7%

Table 3 Performance metrics of models using ISOMAP method

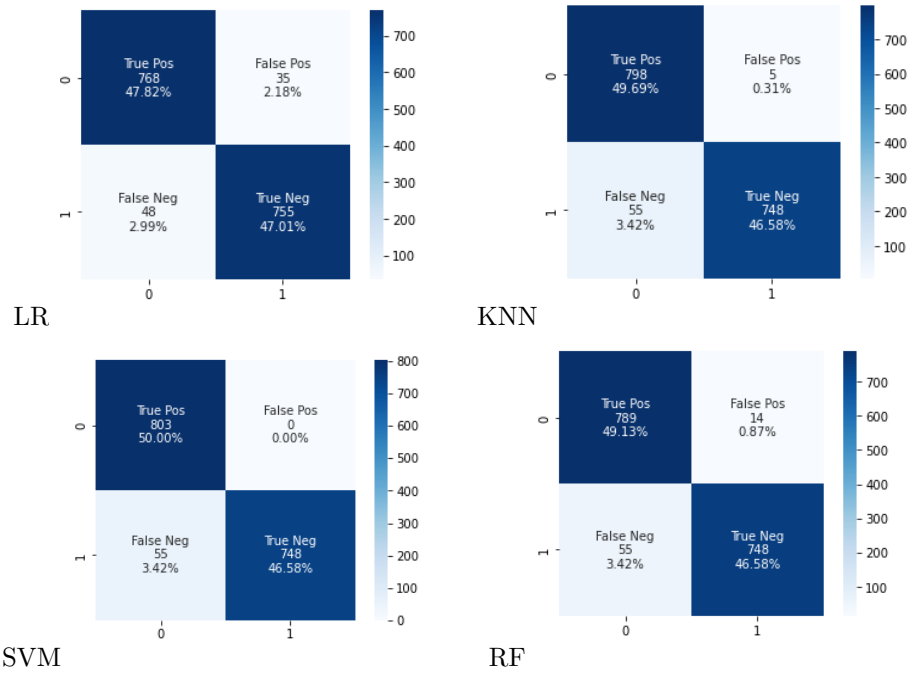

Fig. 16 Confusion matrices of models using LLE method

5.3 Evaluation of LLE

5.3.1 Confusion matrices

The confusion matrices of LLE is given in (Figure 16)

5.3.2 Metrics

The performance metrics of model using LLE is given in (Table 4)

Metric	LR	KNN	SVM	RF
Accuracy	94.8%	96.2%	96.5%	95.7%
Precision	95.5%	99.3%	100%	98.1%
Recall	94.0%	93.1%	93.15%	93.1%
F1 score	94.7%	96.1%	96.4%	95.5%

Table 4 Performance metrics of models using LLE method

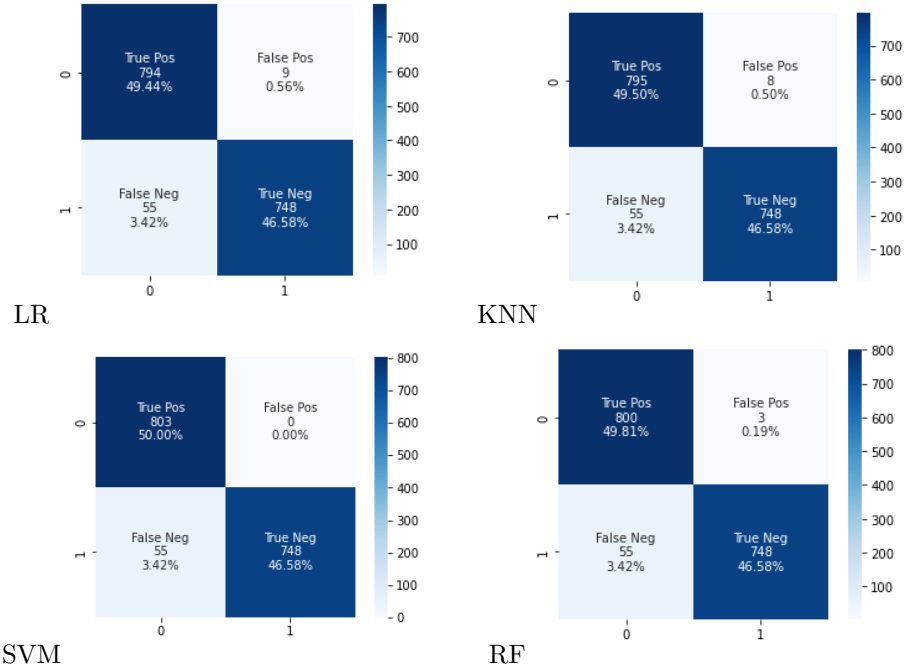

Fig. 17 Confusion matrices of models using MDS method

5.4 Evaluation of MDS

5.4.1 Confusion matrices

The confusion matrices of MDS is given in (Figure 17)

5.4.2 Metrics

The performance metrics of model using MDS is given in (Table 5)

Metric	LR	KNN	SVM	RF
Accuracy	96.01%	96.0%	96.5%	96.3%
Precision	98.8%	98.9%	100%	99.6%
Recall	93.1%	93.1%	93.15%	93.1%
F1 score	95.8%	95.9%	96.4%	96.2%

Table 5 Performance metrics of models using MDS method

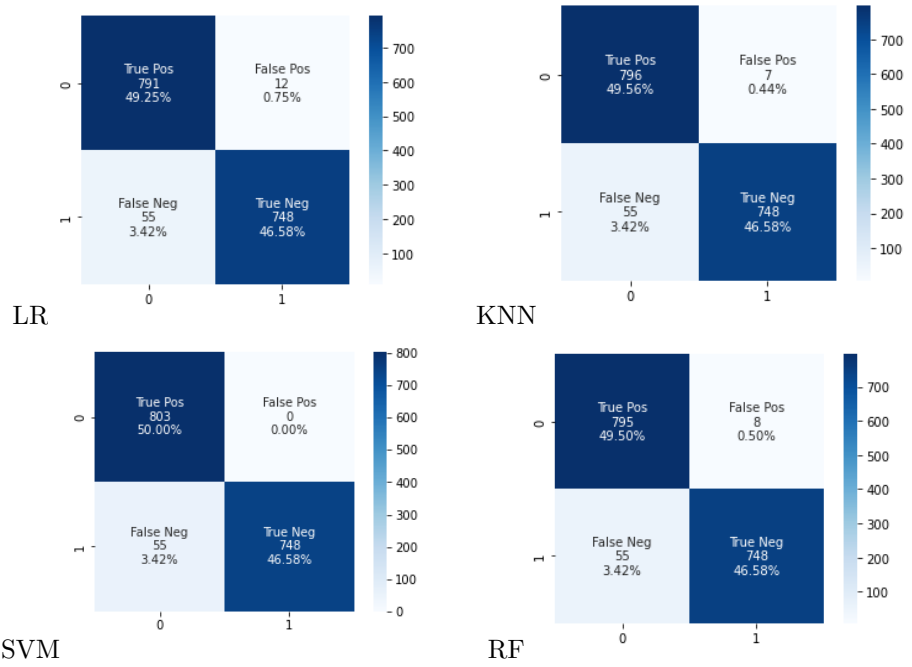

Fig. 18 Confusion matrices of models using T-SNE method

5.5 Evaluation of T-SNE

5.5.1 Confusion matrices

The confusion matrices of T-SNE is given in (Figure 18)

5.5.2 Metrics

The performance metrics of model using T-SNE is given in (Table 6)

Metric	LR	KNN	SVM	RF
Accuracy	95.8%	96.1%	96.5%	96.0%
Precision	98.4%	99.0%	100%	98.9%
Recall	93.1%	93.1%	93.1%	93.1%
F1 score	95.7%	96.0%	96.4%	95.9%

Table 6 Performance metrics of models using T-SNE method

5.6 Comparison with other research paper

Research paper	Accuracy
Paper [2]	95.3%
Paper [3]	92.6%
Paper [4]	87.5%
Our result	96.2%

Table 7 Performance analysis of other research paper compared to our proposal for KNN classification model

Research paper	Accuracy
Paper [2]	-%
Paper [3]	-%
Paper [4]	88.19%
Our result	96.5%

Table 8 Performance analysis of other research paper compared to our proposal for SVM classification model

Research paper	Accuracy
Paper [2]	-%
Paper [3]	76.9%
Paper [4]	49.06%
Our result	96.5%

Table 9 Performance analysis of other research paper compared to our proposal for LR classification model

Research paper	Accuracy
Paper [2]	87.90%
Paper [3]	96.4%
Paper [4]	-%
Our result	96.3%

Table 10 Performance analysis of other research paper compared to our proposal for RF classification model

6 Conclusion and Future work

This research presented an analysis of non-linear dimensionality reduction techniques such as ISOMAP, MDS, T-SNE, LLE on basic classification models such as KNN, LR, SVM, RF on cervical cancer dataset. The data has undergone numerous data preprocessing like replacing null values, class imbalance handling problem. After training and evaluating models, we can conclude that accuracy of all classification models using MDS is highest as compared to others. While the accuracy of LR has substantially increased using dimensionality reduction technique. In the future, one can do analysis of all the tests i.e Hinselmann, Schiller, Cytology using different classification models.

References

[1] Nour Nawal M: Cervical cancer: a preventable death. PMC2812875 **2**, 240–4 (2009)

[2] Dhwaani Parikh, Vineet Menon: Machine learning applied to cervical cancer data. International Journal of Mathematical Sciences and Computing(IJMSC) **Vol.5**(No.1), 53–64 (2019). https://doi.org/10.5815/ijmsc.2019.01.05

[3] Nazim Razali, Salama A Mostafa, Aida Mustapha, Mohd Helmy Abd Wahab, Nurul Atieqah Ibrahim: Risk factors of cervical cancer using classification in data mining. Journal of Physics: Conference Series (2019). https://doi.org/10.1088/1742-6596/1529/2/022102

[4] Choudhury, Avishek: Classification of cervical cancer dataset. IISE ANNUAL CONFERENCE 2018 (2018)

[5] Kaushalya Dissanayake, Md Gapar Md Johar: Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing (2021). https://doi.org/10.1155/2021/5581806

[6] Roopa, H. and Asha, T.: A linear model based on principal component analysis for disease prediction. IEEE Access (2019). https://doi.org/10.1109/ACCESS.2019.2931956

[7] Andreas Buja and Deborah F. SWAYNE and Michael L. Littman and Nathaniel Dean and Heike Hofmann: Xgvis: Interactive data visualization with multidimensional scaling (2001)

[8] Laurens van der Maaten: Accelerating t-sne using tree-based algorithms. Journal of Machine Learning Research (2014)

[9] Yousaf, M., Rehman, T.U. Jing, L.: An extended isomap approach for nonlinear dimension reduction. SN Computer Science (2020). https://doi.org/10.1007/s42979-020-00179-y

[10] Saul,Lawrence and Roweis Sam: An introduction to locally linear embedding. Journal of Machine Learning Research (2001)

[11] Fernandes, K., Cardoso, J.S., Fernandes, J.: Transfer learning with partial observability applied to cervical cancer screening, 243–250 (2017)