
Proceedings Paper:
Fabiyi, SD orcid.org/0000-0001-9571-2964 and Ezechukwu, DN (2022) Feature Extraction and Dimensionality Reduction of Cancer Data Using Folded LDA. In: 2022 3rd International Informatics and Software Engineering Conference (IISEC). 3rd International Informatics and Software Engineering Conference, 15-16 Dec 2022, Ankara, Turkey. IEEE . ISBN 978-1-6654-5996-9

https://doi.org/10.1109/IISEC56263.2022.9998312

Feature Extraction and Dimensionality Reduction of Cancer Data Using Folded LDA

Samson Damilola Fabiyi
School of Computing
University of Leeds
Leeds, United Kingdom
s.d.fabiyi@leeds.ac.uk

Dismas Ndubuisi Ezechukwu
Software and Verification Department
Stoneridge Electronics AS
Saku vald, Harju, Estonia
Ezechukwu.DismasNdubuisi@stoneridge.com

***Abstract*—Linear Discriminant Analysis is a less commonly applied dimensionality reduction technique in cancer data classification. This could be due to the inability of LDA to achieve good classification results when applied on small training data – a common characteristics of cancer data. F-LDA is an extension of LDA and was recently proposed in another application to overcome the challenge posed by the lack of enough samples for training. This paper therefore evaluates the effectiveness of F-LDA as a dimensionality reduction technique in cancer data classification. Experimental results obtained are promising and demonstrate the ability of F-LDA to effectively reduce the dimensionality of cancer data in small training sample scenarios.**

Keywords—F-LDA, LDA, cancer, classification, feature extraction, dimensionality reduction

I. INTRODUCTION

Cancer, when not detected early, can be difficult to manage [1, 2]. Detection of cancer is traditionally performed by medical practitioners. Such practitioners are humans and so are prone to making mistakes [3]. Machine learning models have become useful in medical fields where they are deployed for computer aided diagnosis [1]. The use of machine learning models in cancer diagnosis ensures that the process is automated and results in increased accuracy.

Features or variables which are present in cancer data are usually fed to machine learning models for cancer data classification. While machine learning models are capable of achieving promising results as demonstrated in related works [2, 4], they suffer from Hughes Phenomenon which limits their performance [5, 6]. Hughes Phenomenon (also known as curse of dimensionality) occurs when the ratio of the number of samples to the number of features is very small [7]. While cancer data are usually not characterised by the presence of too many features, they contain very small number of samples due to high cost of labelling and rareness of some diseases [8]. This results in small ratio of sample size to number of features thereby limiting the classification ability of machine learning models [1, 9].

By applying feature extraction techniques on cancer data to reduce its dimensionality, the challenge posed by Hughes phenomenon can be overcome and consequently, performance of classification models are enhanced. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [6] are examples of such techniques. Table I present a summary of different dimensionality techniques which have been applied in cancer data classification. As can be seen in Table I, PCA is a more commonly applied dimensionality techniques than LDA on cancer data. This lack of preference for LDA, as can be seen in Table I, may not be unconnected to the inability of LDA to achieve good results when applied on small training datasets [10]. Recently, an extension of LDA, named F-LDA (F-LDA) [11] was developed to overcome this problem. This paper therefore evaluates the effectiveness of F-LDA as a dimensionality reduction technique in cancer data classification.

II. METHODS AND MATERIALS

A. Datasets Description

The first dataset used in this work is the Breast Cancer Wisconsin (Diagnostic) dataset [12] which contains 569 samples and two classes, namely malignant and benign. There are 30 features in the breast cancer datasets. Another data which is used in this work is the Prostate Cancer Dataset [13] which contains 100 samples and two classes, namely malignant and benign. There are 8 features in the prostate cancer datasets. To create a small sample size classification scenarios, 7 samples are selected from each of the two cancer datasets to form the training sets while the rest are used for testing.

B. Feature Extraction and Dimensionality Reduction Using An Extension of LDA, named Folded LDA (F-LDA)

LDA is a statistical technique which is applied to reduce the dimensions of data. LDA reduces the dimensions of data by maximizing the between-class variance and minimizing the within-class variance in the data. LDA makes use of the labels or classes in the datasets to achieve this goal, hence it is a supervised dimensionality reduction technique.

F-LDA converts each feature vector in the data into a 2D matrix. The resulting data can be considered as a pile of matrices, with each matrix in the data depicting a sample (a 2D sample). F-LDA then apply the conventional LDA steps (within-class matrix computation, between-class matrix computation and data projection) on the resulting data. F-LDA concludes the dimensionality reduction process by unfolding the projected samples (matrix-vector conversion of projected samples). The unfolded samples can then be fed to relevant machine learning models for classification.

C. Classification

SVM is the machine learning model used in this work to classify the breast and prostate cancer datasets. SVM classify the data by using kernel functions (such as the Radial Bias

978-1-6654-5995-2/22/$31.00 ©2022 IEEE

TABLE I. A SURVEY ON DIMENSIONALITY REDUCTION TECHNIQUES (DRT) FOR CANCER DATA CLASSIFICATION

References	DRT	Cancer Types	Classifiers	Year
[14]	PCA	SRBCT, HGG and Lung	Brain Emotional Learning (BEL)	2014
[15]	PCA	Colon, Prostate, Leukaemia, and Lung	Multi-layer Perceptron (MLP)	2020
[16]	Wavelets (WT)	breast cancer	IT2FLS-KMIP IT2FLS-GCCD	2015
[17]	EM-PCA	Breast Cancer	CART-Fuzzy Rule-based	2017
[18]	PCA	Breast Cancer	Naïve Bayes, SVM	2017
[19]	PCA	Breast Cancer	SVM	2020
[20]	PCA, LDA	Breast Cancer	Neural Network, Deep Neural network	2021
[21]	PC+TFS	Breast Cancer	Logistic Regression(LR)	2021
[22]	FS+RP	Breast Cancer	SVM	2016

Function) to find an optimal hyperplane in higher dimensional space. The SVM model is trained in this work using 10 fold cross-validation. The experiment is repeated 10 times and the average classification results reported and presented in the next section. The SVM classifier, and the dimensionality reduction approaches used in this work are implemented in MATLAB.

III. RESULTS AND DISCUSSION

A. Classification Using Breast Cancer Dataset

Firstly, the breast cancer dataset is used to train the SVM classifier. Outputs of LDA and F-LDA, when applied on the breast cancer dataset, are then separately used to train the classifier. When using the LDA, the number of extracted features is varied starting from 1 up to to $c-1$ where c is the number of classes in the dataset. For F-LDA, different configurations (dimensions) of the converted matrices are empirically exploited and the one that gives the best classification result is selected as the optimal parameter.

In other to compare performance of F-LDA with other techniques which are commonly used to overcome the problem of small sample size, Nonparametric Weighted Feature Extraction (NWFE) [23] and Generalized Discriminant Analysis (GDA) [24] are applied separately on the breast cancer dataset. Gaussian kernel is selected as the GDA's kernel function and its parameter (width) optimized in the range $[10^1, 10^2,\ldots, 10^3]$. Finally, PCA, a commonly used dimensionality reduction tool in cancer data classification, is applied. When using the PCA, the number of selected principal component is varied starting from 1 up to 5. The resulting features from the aforementioned processes are used to train the classifier.

The classification results are presented in Table II. As can be seen in Table II, the OA (Overall Accuracy) and AA (Average Accuracy) achieved by LDA are lower than those achieved using the original features. This is expected as LDA is known for giving below par performance when used in small sample size scenarios [10]. Also, as shown in Table II, though OA and AA achieved by NWFE and GDA are improvements on the performance of the LDA, they are still lower than those achieved using the original features. PCA can be seen in Table II to achieve OA and AA which are higher than those given by the original features. Finally, as can be seen in Table II, F-LDA achieved the highest OA and AA. The results achieved by F-LDA on the breast cancer datasets are promising and therefore demonstrate its potentials in breast cancer classification.

B. Classification Using Prostate Cancer Dataset

The operations performed on the breast cancer dataset in Section A above are repeated on the prostate cancer dataset and the classification results presented in Table III. As shown in Table III, outputs of LDA (when applied on the prostate cancer dataset) achieved lower OA and AA than the original features.

TABLE II. CLASSIFICATION RESULTS USING THE BREAST CANCER DATASET

Features	OA (%)	AA (%)
Original	82.40 ± 8.89	77.97 ± 12.64
LDA	63.93 ± 9.31	59.53 ± 10.46
F-LDA	**86.39 ± 5.79**	**82.90 ± 8.37**
NWFE	81.65 ± 8.37	77.42 ± 12.72
GDA	76.71 ± 7.16	70.52 ± 8.61
PCA	83.97 ± 6.71	81.44 ± 8.26

TABLE III. CLASSIFICATION RESULTS USING THE PROSTATE CANCER DATASET

Features	OA (%)	AA (%)
Original	68.92 ± 9.09	64.46 ± 11.29
LDA	56.67 ± 11.28	58.28 ± 8.73
F-LDA	**74.62 ± 7.10**	**72.94 ± 8.69**
NWFE	72.90 ± 10.58	68.00 ± 13.92
GDA	58.17 ± 9.19	56.05 ± 9.32
PCA	72.04 ± 4.15	68.03 ± 5.91

Again, this is expected as LDA is known for giving below par performance when used in small sample size scenarios [10]. As can be seen in Table III, NWFE improves the classifier performance when compared to those given by the original features while GDA continues to achieve lower OA and AA. PCA continues to achieve OA and AA which are higher than those given by the original features, as can be seen in Table III. Finally, as can be seen in Table III, the highest OA and AA are achieved by F-LDA. The results achieved by F-LDA on the prostate cancer dataset are promising and once again demonstrate its potentials in prostate cancer classification.

IV. CONCLUSION

Effectiveness of F-LDA as a dimensionality reduction technique in cancer data classification has been evaluated. Using the breast cancer and prostate cancer datasets, experimental results show that F-LDA can effectively reduce cancer data dimensionality and produce promising classification results in small sample scenarios.

REFERENCES

[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, Jan. 2015.

[2] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer," Ann. Med. Surg., vol. 62, p. 53, Feb. 2021.

[3] A. Mert, N. Kiliç, E. Bilgili, and A. Akan, "Breast cancer detection with reduced feature set," Comput. Math. Methods Med., vol. 2015, 2015.

[4] B. Arivuselvam, S. Tanisha, S. Shalini, and V. S. Subhalaksmi, "Skin Cancer Detection And Classification Using Svm Classifier," Turkish J. Comput. Math. Educ. Res. Artic., vol. 12, no. 13, pp. 1863–1871, 2021.

[5] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[6] S. D. Fabiyi et al., "Comparative Study of PCA and LDA for Rice Seeds Quality Inspection," in 2019 IEEE AFRICON, 2019, pp. 1–4.

[7] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 5, pp. 2297–2307, May 2010.

[8] M. Nagy, N. Radakovich, and A. Nazha, "Machine Learning in Oncology: What Should Clinicians Know?," JCO Clin. Cancer Informatics, no. 4, pp. 799–810, Nov. 2020.

[9] Y. Xiong, M. Ye, and C. Wu, "Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble," Comput. Math. Methods Med., vol. 2021, 2021.

[10] W. Liao, A. Pižurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," IEEE Trans. Geosci. Remote Sens., vol. 51, no. 1, pp. 184–198, 2013.

[11] S. D. Fabiyi, P. Murray, J. Zabalza, and J. Ren, "Folded LDA: Extending the Linear Discriminant Analysis Algorithm for Feature Extraction and Data Reduction in Hyperspectral Remote Sensing," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 12312–12331, 2021.

[12] Breast Cancer Wisconsin (Diagnostic) Data Set, UCI Machine Learning Repository, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

[13] Prostate Cancer, Kaggle, [Online]. Available: https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer?group=bookmarked.

[14] E. Lotfi and A. Keshavarz, "Gene expression microarray classification using PCA-BEL," Comput. Biol. Med., vol. 54, pp. 180–187, Nov. 2014.

[15] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," CAAI Trans. Intell. Technol., vol. 5, no. 1, pp. 22–33, 2020.

[16] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Medical data classification using interval type-2 fuzzy logic system and wavelets," Appl. Soft Comput. J., vol. 30, no. July 2019, pp. 812–822, 2015.

[17] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," Telemat. Informatics, vol. 34, no. 4, pp. 133–144, Jul. 2017.

[18] S. N. Manoli and P. S.K, "Study and Analysis of Breast Cancer Data," Int. J. Eng. Res. Technol., vol. 5, no. 21, Apr. 2018.

[19] W. Wu and S. Faisal, "A data-driven principal component analysis-support vector machine approach for breast cancer diagnosis: Comparison and application," Trans. Inst. Meas. Control, vol. 42, no. 7, pp. 1301–1312, Apr. 2020.

[20] A. Joshi and A. Mehta, "Breast cancer data classification using neural network and deep neural network techniques," Int. J. Recent Sci. Res., vol. 9, no. 4, pp. 25788–25792, 2018.

[21] P. Dhivya, A. Bazilabanu, and T. Ponniah, “Machine Learning Model for Breast Cancer Data Analysis Using Triplet Feature Selection Algorithm,” https://doi.org/10.1080/03772063.2021.1963861, 2021.

[22] H. Xie, J. Li, Q. Zhang, and Y. Wang, “Comparison among dimensionality reduction techniques based on Random Projection for cancer classification,” Comput. Biol. Chem., vol. 65, pp. 165–172, Dec. 2016.

[23] B. C. Kuo and D. A. Landgrebe, “Nonparametric weighted feature extraction for classification,” IEEE Trans. Geosci. Remote Sens., vol. 42, no. 5, pp. 1096–1105, May 2004.

[24] G. Yang, X. Yu, and X. Zhou, “Hyperspectral image feature extraction based on generalized discriminant analysis,” in XXIst ISPRS Congress, 2008, pp. 285–290.