PAPER 125 – ANALYSING GENE EXPRESSION DATA OF PATIENTS WITH AND WITHOUT OVARIAN CANCER USING DYNAMIC MODE DECOMPOSITION

A. M. Usman[1,3*], N. T. Surajudeen-Bakinde[1], F. O. Ehiagwina[1,4], A. S. Afolabi[1], A. A. Oloyede[2], O. S. Zakariya[1]

[1] Department of Electrical & Electronics Engineering, University of Ilorin, Ilorin, Nigeria.
[2] Department of Telecommunication Science, University of Ilorin, Ilorin, Nigeria.
[3]Department of Electrical & Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa.
[4]Department of Electrical & Electronic Engineering, The Federal Polytechnic, Offa, Nigeria.
[*]Email: usman.am@unilorin.edu.ng

ABSTRACT
Machine learning (ML) algorithms have been deployed in recent years as models for the analysis of complex data. The ubiquity of ML algorithms stems from their ability to learn the patterns and structures inherent in data. In addition, they are adaptable to a wide range of data types, irrespective of size, which allows them to learn and predict the future pattern of data. In this work, dynamic mode decomposition (DMD), an ML algorithm, is deployed to analyse the pattern of gene expression data from patients with and without ovarian cancer. Ovarian cancer is one of the deadliest gynaecologic cancers in the world. The obscure nature of the symptoms makes early detection of ovarian cancer difficult. If the disease is diagnosed early, the chance of survival increases for patients. In this work, DMD is applied to analyse gene expression data from patients with and without ovarian cancer to understand the spatiotemporal patterns of the data. The DMD modes captured the prevalent structures and predicted the future state of the data. The results obtained show that DMD is a promising algorithm that can predict the features inherent in gene expression for patients with and without ovarian cancer. The DMD modes can further be applied as features to train detection and classification models that can assist health practitioners in the quest for early detection of ovarian cancer through gene expression data.

KEYWORDS: DMD, eigendecomposition, eigenvalues, eigenvectors, ML, ovarian cancer, SVD

28. INTRODUCTION

Ovarian cancer, a disease often known as a silent killer, is the deadliest gynaecologic malignancy worldwide. It is frequently not diagnosed at the early stage until it has reached an advanced stage because of its generally elusive symptoms. Thus, making it challenging to treat on a medicinal basis (Stewart et al., 2019). Early detection of ovarian cancer in patients increases the chances of survival to 90% at stage 1, 70% at stage 2, and 20% or less survival rate for stages 3-4. Unfortunately, only 20% of ovarian cancers are detected in stage 1 or stage 2 (Elias et al., 2018). In order to achieve accurate early detection of the disease, detection and classification models must show high sensitivity and accuracy in performance, because of the cryptic nature of the symptoms of the disease at the early stages (Elias et al., 2018; Stewart et al., 2019).

Machine learning (ML) algorithms have greatly transformed the way complex data are analysed in recent years. ML algorithms have been deployed in recent years as models for the analysis of complex data. The ubiquity of ML algorithms stems from their ability to learn the patterns and structures inherent in data that may not be feasible for humans to learn. Besides, ML algorithms are applicable to a wide range of data types, regardless of the dimension of the data, whether structured, unstructured, or semi-structured. Interestingly, ML algorithms are also flexible and easily adaptive to variations in data, thereby allowing them to continuously learn the patterns as they evolve and improve prediction along the way. ML algorithms have been used to solve complex problems in different fields of applications which include but are not limited to weather forecasting, marine mammal species detection and classification, financial market analyses, and disease modelling among others (Brunton et al., 2016; Kutz et al., 2016; Usman et al., 2020).

ML algorithms have been deployed for the analysis of different cancer-related datasets. Akazawa & Hashimoto, (2020) deployed five ML algorithms (support vector machines (SVM), random forest, naive Bayes, logistic regression, and XGBoost) to predict the pathological diagnosis of ovarian tumours using patient information and data from preoperative examinations. Diagnostic results were derived from 16 features that were commonly available from blood tests, patient background, and imaging tests. XGBoost provided the best accuracy of 80% among the other classifiers. Prabhakar & Lee, (2020) proposed an ML-based model for classifying ovarian cancer. To choose the most pertinent features for classification, the authors used a variety of datasets, including gene expression data and clinical data, and employed a stochastic optimisation technique. Different feature selection techniques were used to select the most important genes among thousands of other sampled genes in order to

avoid computational complexity. The selected features were passed into different classifiers, and the combination of the genetic bee colony optimisation (GBCO) feature selection technique and the SVM-radial basis function (RBF) kernel technique classifier gave the highest classification accuracy of 99.48%. Wang et al., (2023) proposed a novel strategy for ovarian cancer detection utilising deep learning methods and second harmonic generation (SHG) imaging. A convolutional neural network (CNN) was used to determine whether ovarian cancer is present or absent from SHG photos as input. A dataset of SHG images of ovarian tissue was used to test the model, and an average classification accuracy of 99.7% was reported.

In this research, dynamic mode decomposition (DMD), an ML algorithm, is deployed to analyse the pattern of gene expression data of patients with or without ovarian cancer. Four thousand gene expression data were collected for each patient. This is a high-dimensional dataset with 4000 likely feature-set per patient. DMD is an entirely data-driven ML algorithm that can analyse the spatiotemporal structure of data and thereafter, make a prediction about the future behaviour of the data. Initially developed for the analysis of experiments and simulations in the field of fluid mechanics, DMD has since been used for the analysis of several complex time-varying datasets (Kutz et al., 2016). Thus, the power of DMD is used to extract the patterns of gene expression in order to use it to predict the future state of the data.

29. DESCRIPTION OF THE ALGORITHM

Dynamic mode decomposition (DMD) is an ML algorithm that is enjoying a wide range of applications in different fields of study. DMD has been deployed for weather forecasting, marine mammal species detection, financial market analysis, and disease modelling (Brunton et al., 2016; Ogundile et al., 2021; Proctor & Eckhoff, 2015; Usman & Versfeld, 2022). DMD is entirely an *equation-free*, data-driven algorithm that does not rely on the mathematical equation governing a system. Rather, DMD relies on raw data to determine the structure that describes the behaviour of the system to be modelled. This property makes DMD attractive for applications with different types of data. DMD is deployed to analyse the pattern of gene expression data of patients with and without ovarian cancer.

The DMD algorithm gives the accurate eigendecomposition of a dataset into the dominant spatiotemporal patterns that best explain the behaviours of the system. The patterns are thus used to construct a model of the process being observed. The data collected are arranged as m snapshots in a large, tall skinny matrix $\boldsymbol{X}$, represented as (Kutz et al., 2016):

$$\boldsymbol{X} = \begin{bmatrix} | & | & \dots & | \\ \boldsymbol{x_1} & \boldsymbol{x_2} & \dots & \boldsymbol{x_m} \\ | & | & \dots & | \end{bmatrix}, \tag{1}$$

where the snapshots, $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_m}$, represent the observations in the dataset and m is the number of snapshots. Each snapshot, $\boldsymbol{x}$, contains n numbers of spatial points saved per time snapshot, with $n \gg m$. The data are then split into two matrices:

$$\boldsymbol{X_1} = \begin{bmatrix} | & | & \dots & | \\ \boldsymbol{x_1} & \boldsymbol{x_2} & \dots & \boldsymbol{x_{m-1}} \\ | & | & \dots & | \end{bmatrix} \in \boldsymbol{X}, \tag{2}$$

$$\boldsymbol{X_2} = \begin{bmatrix} | & | & \dots & | \\ \boldsymbol{x_2} & \boldsymbol{x_3} & \dots & \boldsymbol{x_m} \\ | & | & \dots & | \end{bmatrix} \in \boldsymbol{X}. \tag{3}$$

$\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are subsets of the dataset as indicated, and $\boldsymbol{X_2}$ is the exact data matrix $\boldsymbol{X}$, only starting from the second column. The DMD algorithm gives a spatiotemporal decomposition of the dataset into a set of dynamic modes. It does this by finding a best-fit operator, ζ, that would produce the leading eigenvalues, ν and eigenvectors, ϕ, of the dataset.

The best-fit operator, ζ, can be expressed in terms of the data matrices as defined by:

$$\boldsymbol{X_2} \approx \zeta \boldsymbol{X_1} \implies \zeta \approx \boldsymbol{X_2 X_1^{\dagger}}, \tag{4}$$

where $\boldsymbol{X_1^{\dagger}}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{X_1}$. DMD efficiently computes the leading eigenvalues, ν and eigenvectors, ϕ of the best-fit operator, ζ. The DMD modes, $\mathbf{M}$, are the eigenvectors of ζ and each mode has a corresponding eigenvalue of ζ.

However, the algorithm avoids the direct computation of ζ, because the data matrices, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are high-dimensional, which makes them inflexible for the computation of ζ. Instead, the dataset is projected onto a low-rank subspace and then an approximate low-dimensional best-fit operator, $\bar{\zeta}$, defined by:

$$\bar{\zeta} = \boldsymbol{X_2 X_1^{\dagger}}, \tag{5}$$

is introduced for the reconstruction of the leading ν and ϕ of ζ, without overtly computing ζ. The cost of the DMD algorithm is the singular value decomposition (SVD), an effective matrix factorisation technique for high-dimensional datasets (Kutz, 2013).

Thus, given the data matrices, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, the DMD modes, **M**, are computed as follows (Tu et al., 2013):

1. The *economy* singular value decomposition (SVD) of $\boldsymbol{X_1}$ is calculated:

$$\boldsymbol{X_1} = \boldsymbol{U} \sum \boldsymbol{V'}, \tag{6}$$

where $\boldsymbol{U} \in \mathbb{R}^{n \times h}$ is the matrix of the left singular vectors of $\boldsymbol{X_1}$, $\Sigma \in \mathbb{R}^{h \times h}$ is a diagonal matrix containing the singular values of $\boldsymbol{X_1}$, and $\boldsymbol{V'} \in \mathbb{R}^{m \times h}$ is the matrix of the right singular vectors of $\boldsymbol{X_1}$. h denotes the rank of the reduced *economy* SVD approximation of $\boldsymbol{X_1}$.

2. The approximate best-fit operator, $\bar{\zeta}$ is calculated:

$$\boldsymbol{X_2} \approx \boldsymbol{X_1}\bar{\zeta} \;\Rightarrow\; \boldsymbol{U} \sum \boldsymbol{V'}\bar{\zeta}, \tag{7}$$

$$\bar{\zeta} = \boldsymbol{U'X_2 V \sum^{-1}}. \tag{8}$$

3. The leading eigenvalues, ν and eigenvectors, ϕ are derived from the computation of the eigendecomposition of $\bar{\zeta}$:

$$\bar{\zeta}\,\phi = \phi\,\nu. \tag{9}$$

4. The modes, **M**, are calculated as defined by:

$$\mathbf{M} = \mathbf{X_2} \boldsymbol{V \sum^{-1}}\,\phi. \tag{10}$$

30. MATERIALS AND EXPERIMENTAL PROCEDURE

a. Data Description

The proposed model is tested on a dataset containing the gene expression data of 216 patients. The gene expression samples are two separate surface-enhanced laser desorption and ionization-mass spectrometry (SELDI-MS) data, one containing proteomic cancerous serums while the other containing normal serums (Conrads et al., 2004). The data was accessed using MathWorks R2021b edition. For each patient, 4000 gene expression data is collected. 121 of the patients are confirmed to have ovarian cancer while 95 are normal patients with no ovarian cancer.

b. Dataset Pre-processing

The data is preprocessed and structured to fit into the DMD architecture. The gene expression for each patient is arranged column-wise as a snapshot and each snapshot contains 4000 samples (gene expressions) representing the number of rows of the data matrix. Consequently, the raw data is represented in the form of a 4000×216 large, tall skinny matrix to represent $\boldsymbol{X}$ in Equation (1). Thus, each column is a snapshot of 4000 gene expressions of a particular patient while the rows represent specific gene expressions at a given spatial point.

c. Performing DMD on the newly formed data matrix *X*

The newly formed 4000×216 data matrix $\boldsymbol{X}$ is used to derive the two observation matrices, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ as defined in Equations (2) and (3). The DMD modes are thus computed as enumerated from Equations (6) to (10). The eigenvalues, ν and the DMD modes, **M** derived from the DMD operation represent the spatiotemporal coherent

patterns in the gene expression data. The ν gives the dynamic characteristics and behaviour of each mode (columns of $\mathbf{M}$), i.e. $\mathbf{M}_\mathrm{i}$ corresponds to eigenvalue, ν_i.

The predicted future solution can be produced for all time in the future using $\mathbf{M}$. The approximate solutions of the data for all future are computed as defined by:

$$\boldsymbol{X_f} = \mathbf{M}\exp(\boldsymbol{\Omega}\mathrm{t_f})\,\boldsymbol{s}, \tag{11}$$

where $\boldsymbol{\Omega} = \log(\nu)\,/(\Delta\mathrm{t_f})$, is the diagonal matrix of the eigenvalues, t_f is the predicted future time, and $\boldsymbol{s} = \mathbf{M}^{\dagger}\boldsymbol{x_1}$, is the initial amplitude of the modes, † is the Moore-Penrose pseudoinverse operator. Equation (11) is the low-rank representation of the original data. The reconstruction of Equation (11) follows Theorem 1 as defined in Tu et al. (2013) which states that each pair of DMD modes, $\mathbf{M}_\mathrm{i}$ and eigenvalue, ν_i correspond to the eigevectors and eigenvalues of ζ, i.e.:

$$\zeta\mathbf{M}_\mathrm{i} = \mathbf{M}_\mathrm{i}\,\nu_i. \tag{12}$$

d. Data Reconstruction from the Modes

Following the decomposition of the data with the DMD operation, the level of correlation in the gene expression of patients with ovarian cancer and those without ovarian cancer was investigated. This was done using the first three $\boldsymbol{V'}$ modes as depicted in Figure 1. As can be observed, Figure 1 demonstrates that the gene data is significantly correlated indicating that many patients have considerable overlap in gene expression.

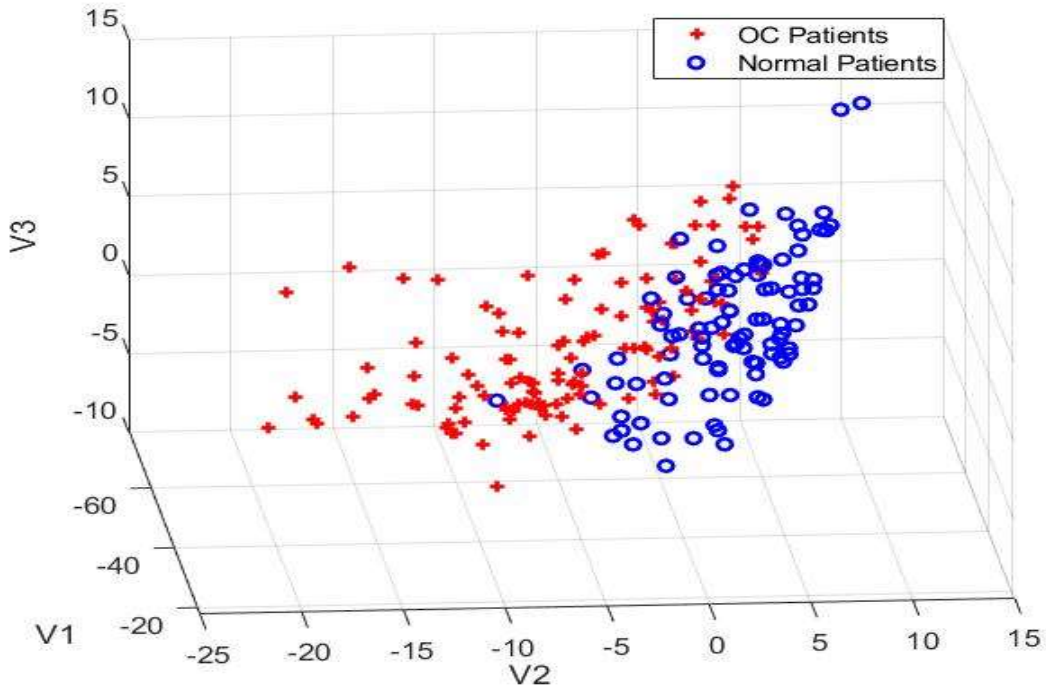

Fig 1: Visualisation of correlation patterns of the gene expression data of patients with and without ovarian cancer in lower-dimensional space.

The computed DMD modes are compared with the gene expression of randomly picked observation (equivalent to the modes positions) to gain an intuition of the accuracy of the computed modes as shown in Figure 2.

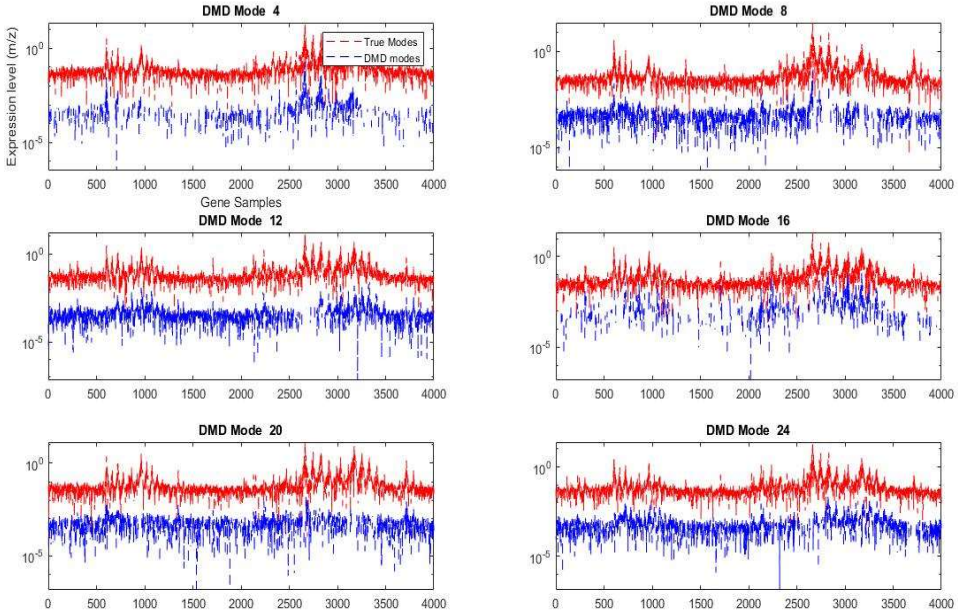

Fig 2: Comparison of the computed modes, **M** with the gene expression samples (True modes).

RESULTS AND DISCUSSION

The rank decomposition level, represented with h, and simulated at different values was implemented. This was done to know the point at which the dominant underlying patterns of the data can be represented with the modes. From the simulated results (at $h = 30$), the computed modes, **M** were randomly picked and compared with the observation snapshots of the original data (true modes) to gain intuition into the accuracy of the DMD. The comparison as depicted in Figure 2 shows that rank-30 decomposition was able to characterise the spatiotemporal pattern of the original data. The predicted data is shown in Figure 3.

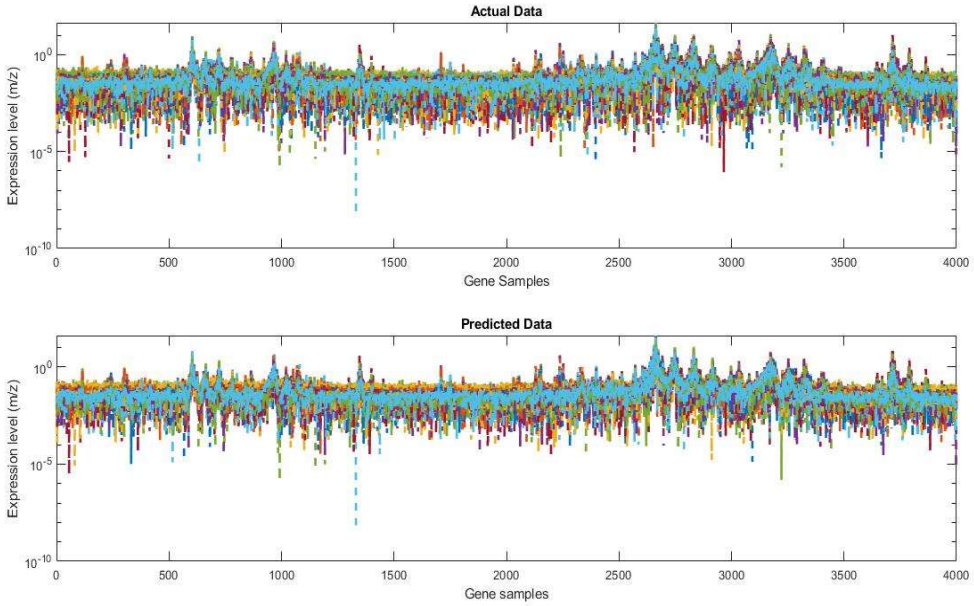

Fig 3: Comparison of the original data and the predicted data (reconstructed using the DMD modes, **M**.

The reconstructed (predicted) patterns demonstrate the effectiveness of DMD in capturing the underlying dynamics that exist in the gene expression data. As can be observed in Figures 2 and 3, the modes and the reconstructed data closely match the original data, thus, indicating that the DMD modes, **M,** capture the majority of the variations that dominate the structure of the original data.

Moreover, the patterns derived from the DMD operation can be interpreted to give insights into the dynamics of the gene expression for patients with and without ovarian cancer. Interestingly, DMD was able to reproduce the dominant structure of the original ovarian cancer data. The ability to reconstruct the original data from the modes provides confidence in the accuracy of the DMD analysis. The benefits of DMD therefore centre on the fact that it is an *equation-free* architecture and that a future-state prediction may be created for any time t (of course, provided the DMD approximation holds).

Therefore, the DMD modes, **M,** can be used as features set to train detection and classification models for accurate detection and classification of gene expressions as either having cancerous cells or not. As earlier noted, the cryptic symptoms of ovarian cancer make early detection difficult. However, ML algorithms can be used to model detection and classifying tools with high performance in terms of sensitivity, accuracy, specificity, and precision. The higher the performance of any of these metrics, the better the chance for early detection of genes that may be carrying the ovarian cancer disease.

Furthermore, it can be observed that despite the original data being of high dimension (4000 x 216), the model can identify patterns and correlations between the gene expression of the patients as depicted in Figure 1 in lower-dimensional space. It is noteworthy that the first three V' modes were plotted in the space encompassed by those modes, patients with ovarian cancer form a discrete cluster distinct from the cluster created by patients without cancer. This shows that high-dimensional data can be effectively analysed in lower-dimensional space. The insights gained from such visualisations can help improve clinical understanding of the disease and its underlying mechanisms.

31. CONCLUSION

The early detection of ovarian cancer among patients could significantly increase the rate of survival. However, detection at the early stage is cryptic due to the elusive nature of the symptoms of ovarian cancer. In this study, DMD was deployed for the analysis of the spatiotemporal coherent patterns that exist in gene expression data of patients with and without ovarian cancer to observe and understand the dominant patterns in the data. An understanding of the patterns can lead to effective feature extractions for training detection and classification models. The DMD modes were used to predict the future state of the data and a closely matched reconstructed pattern of the original data was achieved. The next phase in the work is to deploy the DMD modes as features to train different detection and classification models for the early detection of ovarian cancer tumours from the gene expression data of patients.

ACKNOWLEDGEMENTS

The Federal Government of Nigeria's NEEDS assessment programme through the University of Ilorin is acknowledged for partly supporting this research. MathWorks R2021b© is acknowledged for the ovarian cancer data deployed in this research. Stellenbosch University is acknowledged for providing access to the MATLAB edition (R2021b) used for the research.

REFERENCES

Akazawa, M., & Hashimoto, K. (2020). Artificial intelligence in ovarian cancer diagnosis. *Anticancer Research*, *40*(8), 4795–4800. https://doi.org/10.21873/anticanres.14482

Brunton, B. W., Johnson, L. A., Ojemann, J. G., & Kutz, J. N. (2016). Extracting spatial--temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of Neuroscience Methods*, *258*, 1–15.

Conrads, T., Fusaro, V. A., Ross, S., Johann, D., Rajapakse, V., Hitt, B. A., Steinberg, S. M., Kohn, E. C., Fishman, D. A., Whitely, G., & And, O. (2004). High-resolution serum proteomic features for ovarian cancer detection. Endocrine-related cancer. *Endocrine-Related Cancer*, *11*(2), 163–178. https://doi.org/10.1677/erc.0.0110163

Elias, K. M., Guo, J., & Bast, R. C. (2018). Early Detection of Ovarian Cancer. *Hematology/Oncology Clinics of North America*, *32*(6), 903–914. https://doi.org/10.1016/j.hoc.2018.07.003

Kutz, J. N. (2013). *Data-driven modeling \& scientific computation: methods for complex systems \& big data*.

Oxford University Press.

Kutz, J. N., Brunton, S. L., Brunton, B. W., & Proctor, J. L. (2016). *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM.

Ogundile, O. O., Usman, A. M., Babalola, O. P., & Versfeld, D. J. J. (2021). Dynamic mode decomposition: A feature extraction technique based hidden Markov model for detection of Mysticetes' vocalisations. *Ecological Informatics*, *63*, 101306.

Prabhakar, S. K., & Lee, S. W. (2020). An Integrated Approach for Ovarian Cancer Classification with the Application of Stochastic Optimization. *IEEE Access*, *8*, 127866–127882. https://doi.org/10.1109/ACCESS.2020.3006154

Proctor, J. L., & Eckhoff, P. A. (2015). Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International Health*, *7*(2), 139–145.

Stewart, C., Ralyea, C., & Lockwood, S. (2019). Ovarian Cancer: An Integrated Review. *Seminars in Oncology Nursing*, *35*(2), 151–156. https://doi.org/10.1016/j.soncn.2019.02.001

Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L., & Kutz, J. N. (2013). On dynamic mode decomposition: theory and applications. *ArXiv Preprint ArXiv:1312.0041*.

Usman, A M, & Versfeld, D. J. J. (2022). Detection of baleen whale species using kernel dynamic mode decomposition-based feature extraction with a hidden Markov model. *Ecological Informatics*, 101766.

Usman, Ayinde M, Ogundile, O. O., & Versfeld, D. J. J. (2020). Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access*, *8*, 105181–105206. https://doi.org/https://doi.org/10.1109/ACCESS.2020.3000477

Wang, G., Zhan, H., Luo, T., Kang, B., Li, X., Xi, G., Liu, Z., & Zhuo, S. (2023). Automated Ovarian Cancer Identification Using End-to-End Deep Learning and Second Harmonic Generation Imaging. *IEEE Journal of Selected Topics in Quantum Electronics*, *29*(4). https://doi.org/10.1109/JSTQE.2022.3228567