Dynamic Mode Decomposition-Based Features for Ovarian Cancer Gene Expression Classification Using Machine Learning

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

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. Additionally, 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 worldwide. The obscure nature of the symptoms makes early detection of ovarian cancer difficult. Early diagnosis increases the chances of survival for patients. Furthermore, the modes computed from DMD are used as features and separately fed into three ML classifiers- support vector machine (SVM), decision tree (DecTr), and *K*-nearest neighbour (KNN) to classify the gene expression into either cancer or non-cancer categories. Multiple metrics- sensitivity, accuracy, precision, and error rate are used to assess the performance of the models, to have a balanced illustration of a model's performance. The DecTr outperforms the other two classifiers- SVM and KNN, across the metrics used in evaluating the performance of the models

KEYWORDS: DMD, eigendecomposition, KNN, machine learning, ovarian cancer, SVD

[Received April. 24, 2024; Revised June 3, 2024; Accepted July 14, 2024] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. 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 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). 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).

This work introduced the use of dynamic mode decomposition (DMD) for the analysis pattern of gene expression data. Gene expression gives intuition into the biochemical activities in tissue and cells, as well as the genomic characteristics of organisms. These attributes make gene expression data a potent source for early detection of cancer. DMD is used to gain insight into the pattern of data to enhance the feature extraction process, which is pivotal to the accuracy and precision of ML algorithm predictions (Alharbi & Vakanski, 2023).

II. RELATED LITERATURE

ML algorithms have been deployed for the analysis of different cancer-related datasets. Akazawa & Hashimoto, (2020) deployed five ML algorithms (support vector machine (SVM), random forest, naive Bayes, logistic regression, and XGBoost) to predict the pathological diagnosis of ovarian tumours using patient information and data from preoperative

*Corresponding author: usman.am@unilorin.edu.ng

doi: http://dx.doi.org/10.4314/njtd.v21i3.2543

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 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. Prabhakar & Lee, (2020) proposed an integrated feature selection approach using standard selection techniques such as correlation coefficients, T-statistics, and Kruskal-Wallis test in combination with ML algorithms to classify ovarian cancer. Taleb *et al*., (2022) proposed two models- SVM, and *K*-nearest neighbour (KNN)- to classify ovarian cancer data. They preprocessed the ovarian cancer data by filling in missing values, removing outliers and data normalisation before dividing the data into training and testing data.

However, some works in the literature failed to identify and extract features from the data before being trained and subsequently tested with the ML algorithms. Features play an important role in the final prediction outputs of ML algorithms (Alharbi & Vakanski, 2023). Thus, it is important to understand the underlying patterns or structures of the data to decide which features to extract and use with the ML algorithms.

This work introduces DMD for the analysis of gene expression data of patients with and without ovarian cancer. DMD is utilised to gain insight into the spatiotemporal patterns of gene expression data. Understanding these patterns leads to a more accurate feature extraction process, which in turn enhances the prediction accuracy and precision of ML algorithms.

Hence, this research contributes to the field of cancer classification using ML algorithms with the introduction of DMD for pattern analysis and feature extraction. The features extracted from DMD can be employed with various ML models for the classification of gene expression data. This research highlights the critical importance of evaluating the performance of ML models using multiple metrics. By employing different metrics, we ensure a more nuanced and informed assessment of model performance, ultimately leading to better decision-making and more robust conclusions. Furthermore, the DMD algorithm presented in this work can be modified by other researchers working on different types of cancer data to enhance the feature extraction process and achieve more accurate model prediction. The *equation-free,* data-driven approach of DMD makes it a versatile algorithm that can be further explored for analysing diverse gene expression data that may pertain to different forms of cancer.

DMD has its roots in the field of fluid mechanics, where it is used to study the spatiotemporal patterns of systems. Four thousand gene expression data were collected for each patient. This is a high-dimensional dataset with 4,000 likely feature sets 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 to use it to predict the future state of the data. Furthermore, the dynamics modes computed from the DMD are used as features for the classification of gene expression data of each patient into either cancer or non-cancer. Three ML models - support vector machine (SVM), decision tree (DecTr), and *K*-nearest neighbour (KNN), are introduced for the classification of the gene expression dataset. The performance of the classifiers is compared using metrics such as sensitivity, accuracy, precision, and error rate.

III. MATERIAL AND METHODS

This section gives a detailed description of the algorithms deployed in this work. First, an overview of the DMD algorithm is presented. Following this, we conduct a comprehensive analysis of the gene expression dataset using DMD. The modes computed from the DMD are subsequently used as features for the classification of the gene expression dataset using three classification models - SVM, DecTr, and KNN.

A. Dynamic Mode Decomposition

Dynamic mode decomposition (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 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; Usman & Versfeld, 2022). In this research, 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):

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

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:

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

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

$\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are taken from $\boldsymbol{X}$, and their column is $m-1$. $\boldsymbol{X_1}$ starts from the first column to the $m-1$ column of $\boldsymbol{X}$, while $\boldsymbol{X_2}$ starts from the second column to the last column of $\boldsymbol{X}$. The objective of DMD is to understand the dynamic behaviour or inherent patterns in the dataset by efficiently capturing the changes between successive snapshots. 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:

$$\boldsymbol{X_2} \approx \zeta \boldsymbol{X_1} \implies \zeta \approx \boldsymbol{X_2 X_1^{\dagger}}, \quad (4)$$

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:

$$\bar{\zeta} = \boldsymbol{X_2 X_1^{\dagger}}, \quad (5)$$

is introduced for the reconstruction of the leading ν and ϕ of ζ, without explicitly 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, $\mathbf{M}$, are computed as follows (Tu *et al*., 2013):

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

$$\boldsymbol{X_1} = \boldsymbol{U} \sum \boldsymbol{V'}, \quad (6)$$

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 $\boldsymbol{X_1}$.

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

$$\boldsymbol{X_2} \approx \boldsymbol{X_1}\bar{\zeta} \implies \boldsymbol{U} \sum \boldsymbol{V'} \bar{\zeta}, \quad (7)$$

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

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

$$\bar{\zeta}\, \phi = \phi\, \nu. \quad (9)$$

IV. The modes, $\mathbf{M}$, are calculated as defined by:

$$\mathbf{M} = \mathbf{X_2} \boldsymbol{V} \sum^{-1} \phi. \quad (10)$$

B. Data Description

The proposed model is tested on a dataset containing the gene expression data of 216 patients. Gene expression can play an important role in the early detection of cancer since it indicates the biochemical activities in tissue and cells besides the genetic traits of an organism. In the dataset used in this work, 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. See Conrads *et al*. (2004) for a detailed description of the dataset. The data was accessed using MathWorks R2021b edition. For each patient, 4,000 gene expression data were collected. 121 of the patients are confirmed to have ovarian cancer while 95 are normal patients with no ovarian cancer.

C. Data Preprocessing

The data was 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 4,000 samples (gene expressions) representing the number of rows of the data matrix. Consequently, the raw data is represented in the form of a $4{,}000 \times 216$ large, tall skinny matrix to represent $\boldsymbol{X}$ in Equation (1). Thus, each column is a snapshot of 4,000 gene expressions of a particular patient while the rows represent specific gene expressions at a given spatial point.

D. Analysing the Gene Expression Data using DMD

The newly formed $4{,}000 \times 216$ data matrix $\boldsymbol{X}$ was used to derive the two observation matrices, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ as defined in Equations (2) and (3). The DMD modes were thus computed as enumerated from Equations (6) to (10). The eigenvalues, ν and the DMD modes, $\mathbf{M}$ 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. $\mathbf{M}_1$ corresponds to eigenvalue, v_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:

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

where $\boldsymbol{\Omega} = \log(v)\ /(\Delta\mathrm{t_f})$, 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 modes positions) to gain an intuition of the accuracy of the computed modes as shown in Figure 1.

.

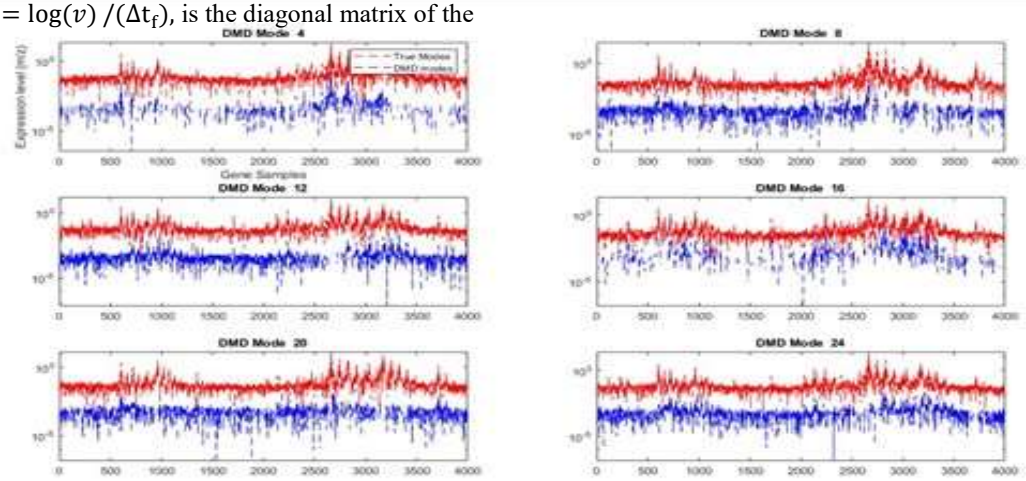

Figure 1: Comparison of the computed modes, M with the gene expression samples (True modes).

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, $\mathbf{M}$ 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 1 shows that rank-30 decomposition was able to characterise the spatiotemporal pattern of the original data. The predicted data is shown in Figure 2.

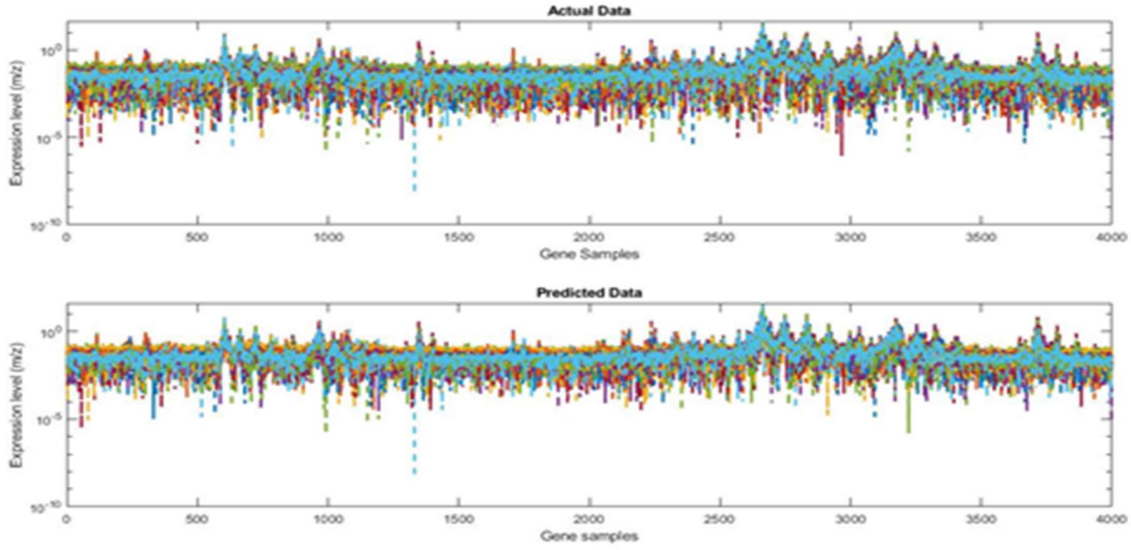

Figure 2: Comparison of the original data and the predicted data (reconstructed using the DMD modes, M.

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 dataset. The ability to reconstruct the original data from the modes provides confidence in the exactness 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).

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 1 and 2, the modes, **M**, and the reconstructed data closely match the original data, thus, indicating that the DMD modes, **M,** captured the majority of the variations that dominate the structure of the original data. Therefore, the DMD modes, **M,** can be used as features set to train ML models for the classification of gene expressions as either having cancerous cells or not.

E *Classification of Gene Expression Data*

The cryptic symptoms of ovarian cancer make early detection difficult. However, ML classifiers are useful tools that are deployed to the detection and classification of cancer data. The feature extraction techniques deployed with the detection and classification models play an important role in determining the performance rates of such models. Feature engineering is a critical stage of the ML process that involves the extraction of features from raw data and translating them into formats suited for ML models. Extracting the proper features will not only make modelling easier but will also result in improved model performance. The technique deployed to extract features from datasets will depend on the structure of the data as well as its adaptability to the ML model in which it will be employed. The DMD modes, **M,** are used as features for three ML classifiers – SVM, Dectr, and KNN to classify the gene expression data into either cancer or normal patients. The schematic diagram of the proposed classification model is shown in Figure 3.

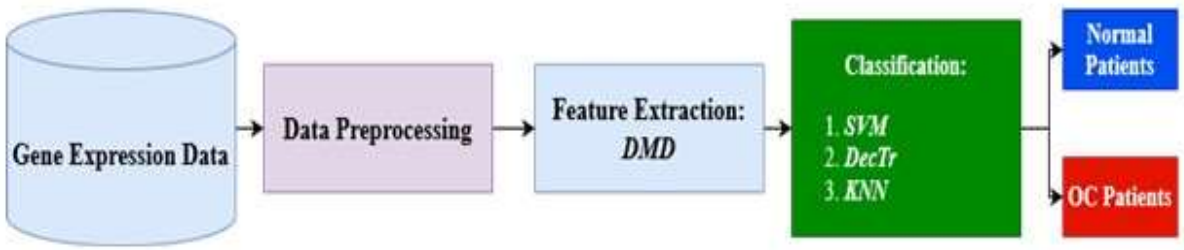

Figure 3: Schematic diagram of the classification model.

I. Support Vector Machine

Support vector machine (SVM) is a supervised ML algorithm fit for classification and regression problems, usually used in solving binary classification problems. The algorithm works by pinpointing the optimal hyperplane that distinguishes data points into two classes (Jakkula, 2006). The hyperplane is localised to split the classes under consideration based on the biggest margin. SVM seeks the greatest margin length while keeping observations in the positive and negative classes apart. SVM has been used in different fields such as natural language processing, signal processing applications, image recognition, prediction, and text categorisation among others. The performance of SVM is highly dependent on parameter selection and settings. The decision function, $\boldsymbol{f(X)}$**,** generates a numerical score or decision value for each instance, and the sign (positive or negative) of that score determines the predicted class. It is mathematically represented as

$$\boldsymbol{f(X) = X'\beta + c}, \tag{12}$$

where $\boldsymbol{X}$ corresponds to an observation, $\boldsymbol{\beta}$ is the vector for coefficients that define the orthogonal vector to the hyperplane, and $\boldsymbol{c}$ is the bias. Jakkula (2006) provides a detailed description of the operating principles of the SVM. The classification is done such that

$$\boldsymbol{f(X) = 0}, \tag{13}$$

represents the hyperplane separation. Therefore, the points above or below the hyperplane separation fulfil the condition $f(X) > 0$ or $f(X) < 0$ to indicate cancer patient or normal patient.

II. Decision Tree

Decision Tree (DecTr) is a hierarchical tree-like, supervised ML algorithm that is used for both classification and regression problems. The algorithm operates by recursively partitioning the input space into territories and assigning a label or value to each territory according to the majority class or average target value of the training instances within that territory.

The major components of DecTr are the tree structure, decision nodes, leaf nodes, splitting criteria, recursive splitting, and predictions. DecTr is used in various fields such as operation research, healthcare, data mining, finance, and marketing among others. It operates in a hierarchical tree-like pattern by generating decisions at each internal node of the tree depending on specified feature values. The decisions drive the path down the tree until it reaches a leaf node, at which point a final prediction is produced. For a feature vector, $\boldsymbol{X}$ and decision rule at a node represented as $\boldsymbol{d}$, the decision rule investigates a specific feature $\boldsymbol{X_i}$ against a threshold $\boldsymbol{g}$. The state of the condition, whether true or otherwise will determine the branch the algorithm will follow, subject to the splitting criteria that are deployed. The interpretability of the outputs of the algorithm is one of its major advantages. A detailed operation of DecTr is discussed in Hernández *et al*. (2021). The DecTr classifier was implemented to effectively partition the features into binary target labels – 'Cancer' and 'Normal' status.

III. K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a supervised ML algorithm that is used also for both classification and regression tasks. The algorithm is an instance-based learning algorithm, which means it remembers the training examples and utilises them directly in the prediction phase. The algorithm's categorisation or assignment of value to a data point is predicted based on the majority class or the average *K*-nearest data point in the feature space. The K represents the number of neighbours that are used in determining the class. Every sample is drawn one at a time from the test set. Then, the Euclidean distance between the entries training set and the test set is computed. Afterwards, the class labels for the training labels are determined. Lastly, the KNN assigns each sample to the class with the most training samples (Cunningham & Delany, 2021). For a feature vector, $\boldsymbol{X}$, 5 is selected as the optimal K value through cross-validation. The Euclidean distance.

$$\boldsymbol{ED(x_i, x_j)} = \sqrt{\sum_{m=1}^{M} (\boldsymbol{x_{im}} - \boldsymbol{x_{jm}})^2}, \qquad (14)$$

of each test data point was computed to all the training data points, where $\boldsymbol{x_{im}}$ and $\boldsymbol{x_{jm}}$ are the m-th features of data points x_i and x_j. The class of the new data point is then predicted by a majority vote of the KNN.

which states that each pair of DMD modes, $\mathbf{M}_i$ and eigenvalue, ν_i correspond to the eigenvectors and eigenvalues of ζ, i.e.:

$$\zeta \mathbf{M}_i = \mathbf{M}_i \, \nu_i. \qquad (15)$$

IV. RESULTS AND DISCUSSION

The $\mathbf{M}$ modes derived from the DMD were fed into each of the classifiers - SVM, DecTr, and KNN, as features. The dataset was divided into two unequal parts, 75% for training data and 25% for testing data. The implementation was carried out on a *Microsoft Windows 10 Pro* computer with an Intel Core (TM) i7-6600U CPU (@2.60GHz) and 16 GB of RAM, running MATLAB 2021b. The performance of the classifiers was evaluated using four metrics, namely- sensitivity, accuracy, precision, and error rate. Each of the metrics relies on the four outcome parameters of the binary classification model. The details of the metrics are explained in Usman, *et al*., (2020) and Usman & Versfeld (2024). The results obtained from the simulation carried out are presented in Table 1.

Table 1: Results of the three classifiers- SVM, DecTr, and KNN.

Classifier	Performance Result (%)			
	Sensitivity	Accuracy	Precision	Error Rate
SVM	76.74	84.49	72.19	15.51
DecTr	89.69	98.93	94.81	1.07
KNN	85.00	90.93	89.98	9.07

The sensitivity, also known as the true positive rate, represents the percentage of correct classifications from each of the models. As can be observed from the results obtained, the DecTr gives the optimal true positive rate, 89.69%, followed by KNN at 85.00%, and SVM at 76.74%. A higher sensitivity output from a classifier provides greater assurance that cancerous gene expression will not be misclassified as normal gene expression. Besides, the sensitivity metric is less affected in a situation where there is a class imbalance in a dataset, as in the case of the dataset used in this research. However, the shortfall of this metric is its potential to give a high number of false positives, i.e. a situation where the model classifies a normal patient as having cancer. This fact buttresses the need to use more than one metric to assess the performance of ML models.

Consequently, the accuracy of each of the classifiers was evaluated. The accuracy metric shows the overall correctness of a model's classification output. From the result, the SVM gives the least accuracy, 84.49%, followed by KNN with 90.93%, while the DecTr is the most accurate of the three classifiers with 98.93%. Accuracy is a comprehensive assessment of the model's performance across all classes (cancer and non-cancer), but it is not effective when there is a class imbalance in the dataset. Therefore, accuracy alone may not provide a comprehensive illustration of a model's performance.

Furthermore, the precision metric was introduced to assess the models' performance. This metric is particularly important in giving a true illustration of the correct classification made by a model. Specifically, precision pays attention to the positive classification done by a model and evaluates how many are truly positive. In the results obtained, the DecTr outperforms the other two classifiers with a 94.81% precision

rate, followed by KNN with 89.98%, and the SVM gives the lowest precision rate of 72.19%. In the context of gene expression data classification, precision helps to assess the true state of a model's classification when it claims that a gene expression is cancerous. It is also a very useful metric to assess a model's performance in situations where the cost of wrong classification is high, such as in a cancer diagnosis, where wrong classification of a normal patient as having cancer could lead to unnecessary treatment and intervention.

The overall comparison of the classifiers as depicted in Figure 4 shows that the DecTr classifier outperforms the other two- SVM and KNN, across the metrics. This can be attributed to the effectiveness of DecTr in handling non-linear relationships between features and the target variable. Besides, the hyperparameter settings played a crucial role in determining the performance output of a classifier. The poor performance of the SVM in comparison with DecTr and KNN can be attributed to SVM's sensitivity to class imbalance in the dataset. Also, DecTr has the lowest error rate among the three classifiers. The error rate is complimentary to the accuracy metric, meaning that a higher error rate corresponds to lower accuracy.

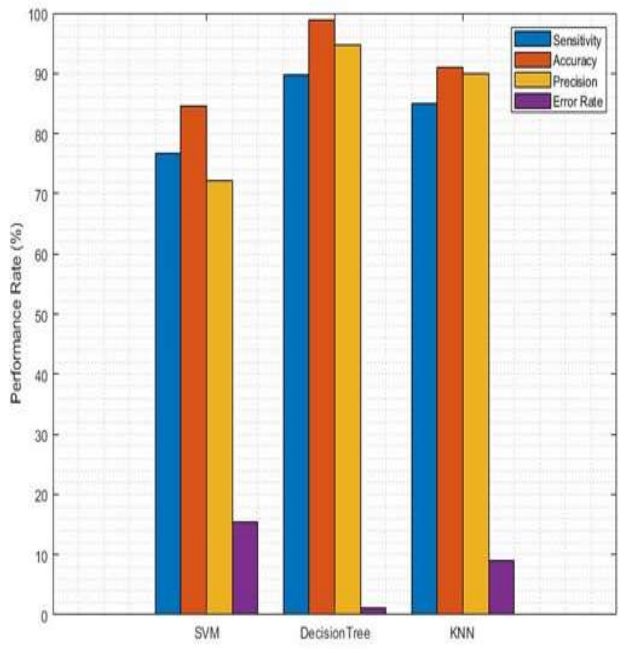

Figure 4: Comparison of the Performance of the three classifiers.

The analyses from the obtained results show the importance of assessing the performance of ML models with more than one metric in order to gain a comprehensive illustration as to which models perform better and under what circumstances, subject to the goals to be achieved. For instance, the sensitivity metric minimises the risk of failing to classify gene expression with cancer, but with the trade-off of wrongly classifying a normal gene expression as having cancer. On the other hand, the accuracy metric shows the overall correctness of a model's performance; however, it may not give a complete picture of the model's performance if there is a class imbalance, as in the case of the data used in this research. Accuracy also fails to distinguish between the two error types (false positive and false negative). In the case of precision, it is very useful in the context of the dataset used in this research because it focuses on the correct classification made by a model. In the case of cancer classification, where the cost of false positives (wrong classification) is high, this is a critical metric to assess models' performance. In addition, precision is also valuable in case of class imbalance in datasets because it is not affected by the number of true negatives (i.e. instances where a model classifies a class as negative and it is truly so). Despite its advantages over the other two metrics, the precision trade-off includes the chance of missing actual cases of cancer classification. Thus, there will always be a trade-off between the metrics. Optimising a model's performance to suit a metric may inadvertently affect another metric. Therefore, the choice of metrics will depend on the specific priorities and trade-offs relevant to the problem at hand.

.

IV. 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. Furthermore, the modes, **M**, computed from DMD were used as features on three ML models – SVM, DecTr, and KNN to classify the gene expression data as either cancer or non-cancer. The DecTr outperformed the other classifiers across the metrics used to assess the performance of the models. This work also buttresses the importance of assessing the performance of ML models with more than one metric in order to have a complete illustration of the working of a model. The evaluation of a model's performance using multiple metrics often gives a more nuanced assessment of the models, subject to the specifics of the goal that is set to be achieved.

AUTHOR CONTRIBUTIONS

A. M. Usman: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing – original draft; Writing - review & editing. **N. T. Surajudeen-Bakinde** Supervision; Writing - review & editing. **F. O. Ehiagwina** Analysis; Writing - review & editing. **A. S. Afolabi** Validation; Writing - review & editing. **A. A. Oloyede** Analysis; Methodology; Writing - review & editing. **O. S. Zakariya** Software; Visualization; Writing - review & editing

ACKNOWLEDGEMENT

The Federal Government of Nigeria's NEEDS assessment programme through the University of Ilorin is acknowledged for partly supporting this research. Stellenbosch University is acknowledged for providing a conducive environment where the research was conducted. MathWorks R2021b© is acknowledged for the ovarian cancer data deployed in this research.

REFERENCES

Akazawa, M. and Hashimoto, K. (2020). Artificial intelligence in ovarian cancer diagnosis. *Anticancer Research*, *40*(8), 4795–4800.

Alharbi, F. and Vakanski, A. (2023). Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering*, *10*(2). 3

Brunton, B. W.; Johnson, L. A.; Ojemann, J. G. and 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 And, O. (2004). High-resolution serum proteomic features for ovarian cancer detection. Endocrine-related cancer. *Endocrine-Related Cancer*, *11*(2), 163–178.

Cunningham, P. and Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. *ACM Computing Surveys*, *54*(6).

Elias, K. M., Guo, J. and Bast, R. C. (2018). Early Detection of Ovarian Cancer. *Hematology/Oncology Clinics of North America*, *32*(6), 903–914.

Hernández, V. A. S.; Monroy, R.; Medina-Pérez, M. A.; Loyola-González, O. and Herrera, F. (2021). A Practical Tutorial for Decision Tree Induction. *ACM Computing Surveys*, *54*(1).

Jakkula, V. (2006). Tutorial on support vector machine (SVM). *School of EECS, Washington State University*, *37*.

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. and Proctor, J. L. (2016). *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM.

Ogundile, O. O.; Usman, A. M.; Babalola, O. P. and 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. and Lee, S. W. (2020). An Integrated Approach for Ovarian Cancer Classification with the Application of Stochastic Optimization. *IEEE Access*, *8*, 127866–127882.

Proctor, J. L. and 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. and Lockwood, S. (2019). Ovarian Cancer: An Integrated Review. *Seminars in Oncology Nursing*, *35*(2), 151–156.

Taleb, N.; Mehmood, S.; Zubair, M.; Naseer, I.; Mago, B. and Nasir, M. U. (2022). Ovary Cancer Diagnosing Empowered with Machine Learning. *2022 International Conference on Business Analytics for Technology and Security, ICBATS 2022*, 1–6.

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

Usman, A. M. and Versfeld, D. J. J. (2024). Principal components-based hidden Markov model for automatic detection of whale vocalisations. *Journal of Marine Systems*, *242* (October 2023), 103941.

Usman, A. M. and 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, A. M.; Ogundile, O. O. and Versfeld, D. J. J. (2020). Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access*, *8*, 105181–105206.

Wang, G.; Zhan, H.; Luo, T.; Kang, B.; Li, X.; Xi, G.; Liu, Z. and 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).