African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

MACHINE LEARNING APPROACH FOR THE PREDICTION OF BLADDER CANCER STAGES BASED ON NEXT-GENERATION SEQUENCING DATA

Imhenkuomon, A.[1], and Omogbhemhe, M. I.[2]
[1]*School of Computer Science & Mathematics, Liverpool John Moores University, Liverpool, England.*
[2]*Department of Computer Science, Ambrose Alli University, Ekpoma, Edo State, Nigeria.*
[1]*brosteey@gmail.com*
[2]*mikeizah@aauekpoma.edu.ng*

ABSTRACT

Purpose: The purpose of this paper is to apply Machine learning algorithms for the classification of various stages of bladder Cancer (BCa) based on RNA-Seq transcriptome per million(TPM) gene counts data and its corresponding pathological stages from the TCGA database. The objective is to assess classification performance across different stages.

Design/Methodology/Approach: This study applied a computational research design on publicly available BCa gene expression data from The Cancer Genome Atlas (TCGA). Multiple supervised machine learning predictive modelling algorithms were trained and evaluated, with a nested cross-validation design. A forward feature selection technique was used to select the best features for ML classifiers, in conjunction with 3-fold nested cross-validation (nCV), applied to binary classification using machine learning algorithms. The dataset preprocessing was carried out in two phases using the R and Python programming languages.

Research Limitation: Reliance on downloaded data raises concerns about the data generator's bias.

Findings: This study suggests that TPM profiles of bulk RNA-seq samples are unreliable for separating adjacent stages of bladder cancer. These findings suggest that bulk transcriptomic data should not be used solely to inform treatment decisions for bladder cancer. Rather, it will be more informative to integrate molecular subtyping with multi-omics data or to make models that can directly predict clinical outcomes.

Practical Implication: In practical terms, these findings suggest that bulk RNAseq TPM transcriptomic data should not be solely relied on for staging bladder cancer in clinical or predictive settings. Instead, more informative approaches such as combining molecular subtypes, integrating multi-omics data, or focusing on models that predict clinical outcomes are likely to provide greater value for decision-making and future research.

Social Implication: This highlights the effect of over-relying on AI diagnostics that do not capture the full biological characteristics, which is essential for protecting patient safety.

Originality/Value: This research examined the application of machine learning algorithms to predict bladder cancer stages using RNA-seq TPM gene-count NGS data from the TCGA database, a method that researchers have not previously considered.

Keywords: *Bioinformatics. bladder cancer. machine learning, next-generation sequencing. RNAseq.*
ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

INTRODUCTION

Cancer is a disease of the genome that grows uncontrollably (Bosserhoff & Kappelmann-Fenzl, 2021). This uncontrolled growth of cells begins at a site in the body and spreads to other parts, a process called cancer metastasis (Kumar et al., 2022). Medically, cancers are classified into stages based on the extent of growth and spread to other organs. Identification of these stages is important for the diagnosis, treatment, and management of patients with cancer.

Over 20 million people are living with cancer today, with 10 million new cases of cancer worldwide and 6 million cancer deaths each year, and according to GLOBOCAN data, an estimated 573,278 people were diagnosed with bladder cancer in 2020, which accounts for roughly 5% of all new cancer diagnoses (Ferlay et al., 2023). This connotes that cancer is the world's leading cause of death. In view of this, machine learning (ML) techniques in cancer research and oncology have recently demonstrated significant improvements in disease diagnosis and detection (Kourou et al., 2021).

High-throughput data, such as those derived from Next Generation Sequencing (NGS) technology and available in the TCGA database, provide adequate and more comprehensive molecular and clinical information for cancer research. Therefore, this study aims to examine the application of ML algorithms for predicting the stages of bladder cancer (BCa) patients using bulk RNA-Seq NGS transcript per million (TPM) gene expression data from the TCGA database. Thus, this might serve as a potential panacea for the diagnosis of BCa patients. The objective of this study is to determine the ability of machine learning classification algorithms to predict bladder cancer (BCa) stages using RNA-seq data from the Cancer Genome Atlas and to evaluate model performance using a robust validation approach in order to provide a more reliable and unbiased assessment of their predictive capability. To achieve these goals, ML classification algorithms, feature selection techniques, the nested cross-validation technique, and performance metrics, such as the confusion matrix, Area under the Curve (AUC), and Receiver Operating Characteristic (ROC), will be applied. Thus, Identification of these stages will assist clinicians/medical doctors in administering treatment and managing cancer patients.

The bladder is a hollow organ located in the lower part of the pelvis. Its main function is to store urine, a liquid waste produced by the kidneys, which is transported from the kidneys to the bladder through tubes called ureters. Figure 1 describes the various stages, grades, and BCa classification with regard to tumour progression. Bladder cancer (BCa), also known as urothelial carcinoma, most commonly arises in the stratified epithelium of the urinary system along papillary and non-papillary pathways (Guo et al., 2020). BCa is the 10th most common cancer, with 3% of world cancer diagnoses, with prevalent statistics in developed countries, and it is more common in men than women (Wigner et al., 2021; Saginala et al., 2020). Among the multiple important risks of BCa include smoking, schistosomiasis infection, and occupational exposure to certain chemicals, with smoking referred to as the most important risk of BCa (Saginala et al., 2020).

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

LITERATURE REVIEW

Since the advent of ML incorporated into computational Biology, there has been a growing trend in cancer research. The field of cancer classification is gaining momentum, driven by state-of-the-art technologies such as NGS and ML algorithms. Currently, various computational methods are employed for cancer subtypes and pathological stage classification, using data sources such as clinical, cancer image, and gene expression data. Recent studies applying machine learning to bladder cancer have used a wide range of data types, each with its own limitations. Many earlier, clinically focused approaches rely primarily on demographic and clinical variables, often based on relatively small sample sizes and with limited validation.

A systematic review by Kourou et al. (2015), along with more recent updates in the field, highlights common methodological concerns, including inadequate preprocessing, lack of external validation, and potential overfitting. Although these approaches have demonstrated some predictive capability, they are often limited in their ability to capture the underlying biological complexity of the disease. Recently, there has been growing interest in machine learning models based on imaging. The study by Zhang et al. (2022) applied deep learning methods to CT imaging to predict outcomes, including survival and treatment response. While these methods show promise, they rely on phenotypic features and do not directly reflect tumour biology, which may limit their interpretability and clinical utility.

In contrast, recent work has demonstrated the value of gene expression data in identifying prognostic signatures and biologically distinct tumour subtypes, rather than attempting to replicate conventional staging systems (Kamoun et al., 2020). This shift reflects a broader recognition that molecular features may not align closely with clinical staging categories. At the same time, advances in sequencing technologies have highlighted important limitations of bulk RNA-seq data.

Issues such as variability introduced by TPM normalisation (Zhao et al., 2020, 2021) and batch effects in large datasets like TCGA (Wang et al., 2018) can affect downstream analyses and model performance. These findings suggest that while machine learning has been widely applied in bladder cancer research, many existing studies are constrained by data limitations, methodological challenges, or a focus on endpoints that may not fully reflect tumour biology. This provides a strong rationale for systematically evaluating the ability of transcriptomic data to predict clinical stage, using a robust, well-validated analytical framework.

Research has shown that BCa is the most expensive cancer to treat, with a total of $4.1 billion yearly on a per-person basis in the United States, with a recurrence rate of 50–80% (Garapati et al., 2017). BCa exists in different forms. Generally, bladder cancer can be divided into muscle-invasive (MIBC) and non-muscle-invasive (NMIBC) based on tumour extent, with approximately 70% of patients diagnosed with NMIBC and the remaining 30% with MIBC or advanced bladder cancer (Kong et al., 2022; Goutas et al., 2021).

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Table 1 provides a brief description of BCa subtypes and describes the clinical-pathological stages, as well as classification based on tumour extent. For efficient treatment of BCa, it is important to ascertain its stages. BCa are staged according to their location and extent of spread to other organs, which are classified as tumour, nodes, and metastases.

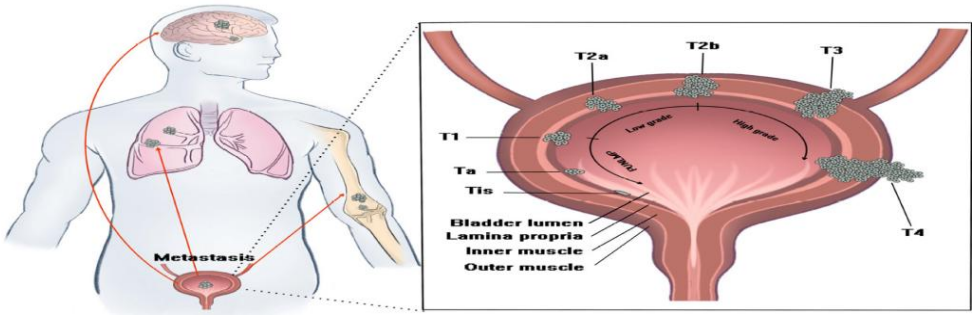

Figure 1: Stages and grades of BCa according to degree of progression (Kong et al., 2022)

Table 1: Bladder cancer stage grouping, subtypes and classification based on Tumour stage

Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Stage Group	Description	Tumour stages	NMIBC & MIBC	Forms of BCa	Description
Stage 0	No primary tumour in the bladder	T0	NMIBC	Transitional cell (urothelial) carcinoma	The most common form of BCa. It makes up approximately 95% of BCa's
Stage 1	Tumor in the connective tissue, but does not involve the bladder wall muscle	T1		Papillary carcinomas	The most common form of BCa. It makes up approximately 95% of BCa's
Stage II	The tumour has spread to the muscle of the bladder wall	T2	MIBC	Squamous cell carcinoma	About 1% - 2% of BCa are squamous cell carcinomas. The majority are Invasive
Stage III	The tumour has spread to the fatty tissue surrounding the bladder	T3		Adenocarcinoma	About 1% of BCa are adenocarcinomas.
Stage IV	The tumour has spread to nearby organs	T4		Flat carcinomas	They do not grow in the hollow part of the bladder. Also referred to as non-invasive flat carcinoma

(*Kong et al., 2022; Garapati et al., 2017)*

Over the past few decades, ML technology has made tremendous improvements across sectors such as healthcare, pharmaceuticals, Economics, Engineering, Agriculture, Life Sciences, and many more. With the advent of Next Generation Sequencing (NGS) technology, ML has been explored in human genetics to predict diseases and their causes (Toh & Brody, 2021).

ISSN: 2408-7920
Copyright ⓒ African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

In the field of Bioinformatics, ML methods have been applied across various biological domains, including proteomics, genomics, systems biology, text mining, and evolution, to extract useful information from biological data (Shastry et al., 2020). The use of ML in clinical decision-making is considered to increase the likelihood of early disease prediction and diagnosis through NGS and high-resolution imaging techniques (Qbal et al., 2021).

Currently, ML methods are being applied across different biomedical areas, ranging from detecting and classifying tumours using X-ray and CT images to classifying malignancies from proteomic and genomic (microarray) assays (Cruz et al., 2006). ML has been used to analyse gene expression data for biomarker identification in cancer research (Zhang et al., 2021). The application of ML in Bioinformatics is gaining momentum, and research is still ongoing.

In the context of cancer research, RNA-Seq gene expression data can serve as indicators of risk of cancer development and progression in tumorous tissue, for classification of cancer types and stages, and for potential response to a specific type of therapy. Next-Generation Sequencing (NGS) is a powerful technology that generates high-throughput data from DNA/RNA molecules. This technology is transforming and making an impact across many fields. For instance, in personalised medicine, clinical diagnostics, cancer research, and genetic diseases are addressed through computational means such as ML. Data generated by NGS technology provides important insights into the functional characteristics of cells and tissues (Bosserhof et al., 2021).

There are several types of NGS technology, and they are categorised based on certain criteria such as existence ("Generation"), sequencing methodology, length of reads, etc. In this research paper, the focus is on Gene Expression Quantification with experimental strategy as "RNA-Seq" gene counts of BCa as warehoused in the TCGA database. The evolution of NGS has enabled the generation of large-scale data, such as transcriptomics. RNA-Seq uses the NGS technology to yield a vast amount of information that continues to fuel discovery and innovation in cancer research and biomedicine in general, including studies on differential gene expression analysis and cancer biomarkers, cancer heterogeneity and evolution, cancer drug resistance, the cancer microenvironment and immunotherapy, and neoantigens (Haque et al., 2017).

RNA-Seq-based analyses have advantages over other classic methods for clinical applications, including precise base-pair resolution, the ability to identify splicing variants, allele-specific expression, novel gene fusions, non-coding RNAs, and novel RNAs (Hong et al., 2020). Therefore, for effective findings in this research paper, the TCGA database has great potential for providing the cancer expression count matrix. However, it is still unclear whether bulk RNA-seq data, particularly TPM-normalised expression profiles, contain sufficient meaningful signal to reliably separate adjacent stages of bladder cancer when evaluated using a more rigorous, unbiased modelling approach.

ISSN: 2408-7920
Copyright Ⓒ African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

METHODOLOGY

This study applied a computational research design on publicly available BCa gene expression data from The Cancer Genome Atlas (TCGA). Multiple supervised machine learning predictive modelling algorithms were trained and evaluated using a nested cross-validation design, with the inner loop dedicated to feature selection and hyperparameter tuning, and the outer loop used to estimate model performance to structure model development and evaluation.

Data Source

The Cancer Genome Atlas (TCGA) Research Network has reported and analysed large numbers of human tumours to discover molecular aberrations at the DNA, RNA, protein, and epigenetic levels (Wang et al., 2020). In this regard, BCa RNA expression data with patients' clinical data were downloaded from TCGA (https://portal.gdc.cancer.gov/) using an R programming package called *TCGAbiolinks* in R version 4.2.1. It was developed as an R/Bioconductor package to address challenges in data mining and analysis of cancer genomics data stored in the Genomic Data Commons (GDC) (Weinstein et al., 2013). Table 2.0 provides an overview of RNA expression components for BCa warehoused in the TCGA database.

Table 2: Transcriptome Profile of Bladder Cancer Data Available in TCGA

	file_count	case_count	data_category
1	2420	412	Copy Number Variation
2	2590	412	Sequencing Reads
3	5858	412	Simple Nucleotide Variation
4	1320	412	DNA Methylation
5	581	412	Clinical
6	1736	412	Transcriptome Profiling
7	1760	412	Biospecimen
8	343	343	Proteome Profiling
9	1723	406	Structural Variation

For the analysis of this data, we require all RNA-Seq data to be queried, as we are more interested in the normalised gene expression count. Hence, *unstranded, stranded_first, stranded_second, tpm_unstrand, fpkm_unstrand, and fpkm_uq_unstrand* RNA-Seq quantification measures were the available assays in the "SummarisedExperiment" class with "tpm_unstrand" as our gene expression matrix, which has 60660 samples and 431 observations. The code snippet below executed our data of interest

```
# querying all RNA-seq data from BLCA project
               query_TCGA = GDCquery(
                   project = "TCGA-BLCA",
                   data.category = "Transcriptome Profiling",
                   data.type = "Gene Expression Quantification",
```

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

```
experimental.strategy = "RNA-Seq",
workflow.type = "STAR - Counts",
barcode = c("TCGA-*")
```

TPM gene normalisation data

As a measure of RNA-Seq quantification, RPKM (reads per kilobase of transcript per million reads mapped) and TPM (transcripts per million) or FPKM (fragments per kilobase of transcript per million reads mapped) all account for sequencing depth and feature-length (Zhao et al., 2021). Research has shown that TPM was proposed due to inaccuracies in RPKM (Zhao et al., 2020).
In this research paper, we did not process raw count data, but we applied TPM count data available in the TCGA database, which was not only ideal, but it was the available data for our research. The TPM expression counts are obtained thus (Zhao et al., 2021):

$$TPM = 10^6 * \frac{Reads\ map\ to\ transcript/transcript\ length}{sum(read\ map\ to\ transcript/transcriptlengt\square)}$$

Data Preprocessing

The preprocessing of the dataset was conducted in two phases using R and Python programming languages. The first phase was executed with R in downloading the datasets of interest as stated in the previous section. To prepare the features (gene expression count matrix) for our ML models, the TPM expression data needs to be processed. In principle, the rows are the gene "Ensemble ID", and the columns are the "sampleID". First, the "ensembleID" was matched with the "gene names" as provided in the "rowData" object of the "summarizedExperiment".

To make the column names the gene names, which are the variables for the prediction, the data frame was transposed (rows became columns and columns became rows). Also, in getting the appropriate clinical metadata, the "casesID" in the gene expression file was identified with the "barcode" in the "colData" (clinical data). In other words, the "casesID" within rows in the gene expression file should be identical to the "barcodeID" in the clinical data file. From the clinical trial data, we are only interested in the AJCC pathological stage as the target/label data. There are four stages of BCa as documented in TCGA patients' clinical data with a total of 431 records before preprocessing, as illustrated in Table 3.

Table 3: AJCC Pathological Stages of BCa Dataset

Stages	Stage1	Stage11	Stage111	Stage1V	Total
No of observation	4	134	148	144	431

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

It was necessary to remove expression counts below a certain threshold because they most likely do not make a significant contribution to our ML model. In this paper, we considered columns with less than 200 gene counts and deleted them. A total of 4931 columns and 428 rows were obtained for ML models.

The methodologies adopted in this research are primarily based on the Python programming language, unlike the initial data extraction and some preprocessing, which were done with R. The figure below illustrates the process flow for this research.

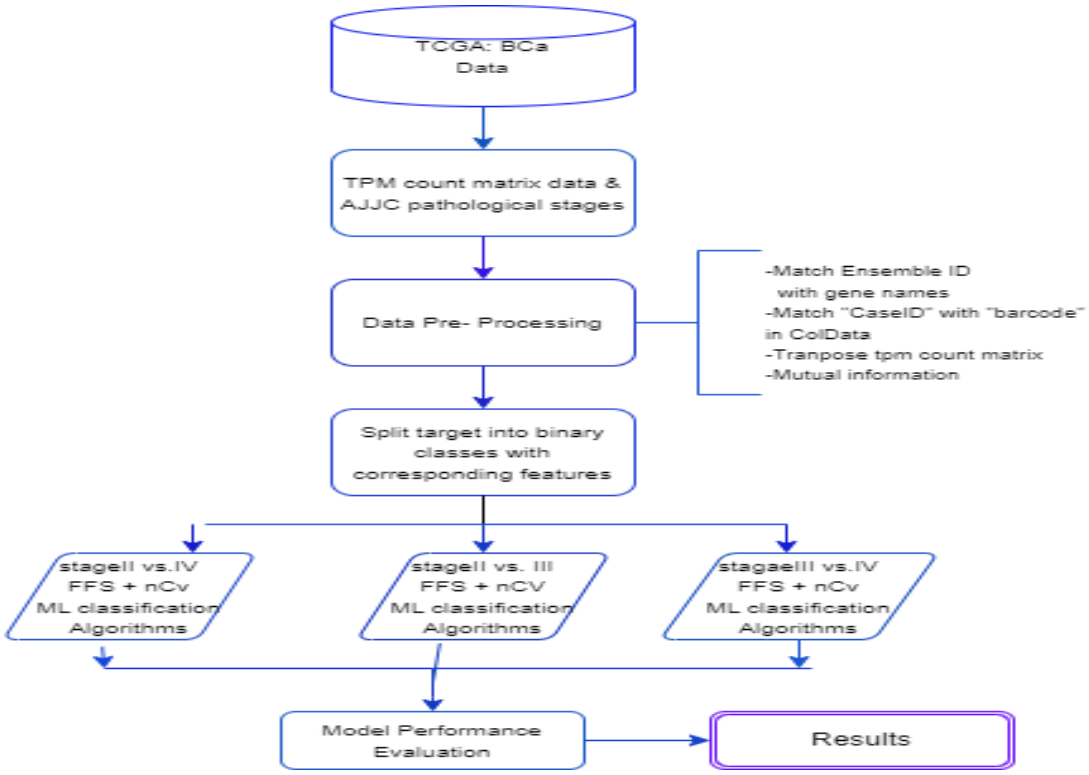

Figure 2: Process Flow of Adopted Methodology

T-distributed Stochastic Neighbour Embedding (t-SNE)

Our initial approach was to gain insight into our data. It is most often important to know the distribution of datasets for ML problems. Computational investigation of high-throughput data usually involves unsupervised exploratory measures such as dimensionality reduction for visualisation. Thus, t-distributed stochastic neighbour embedding (t-SNE) was used. This

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

approach maps a set of high-dimensional points to two dimensions; by that, close neighbours remain in proximity and distant points remain distant, and the algorithm places all points on the 2D plane, first at random positions to interact in a manner that assumes the points to be like physical particles (Kobak & Berens, 2019).

A clearer picture between the features (genes) and the BCa patients' stages (Stage I, Stage II, Stage III, and Stage IV) was examined. This technique was adopted for its ability to capture nonlinear relationships in big data, such as RNA-seq, compared with Principal Component Analysis (PCA), nonnegative matrix factorisation (NMF), and classical multidimensional scaling (MDS) (Xu et al., 2020). Two components were applied in our work with this technique as a parameter for setting up the visualisation using the "manifold" package from Scikit-learn in Python.

Mutual Information

Dimensionality reduction/feature selection are fundamental preprocessing steps for analysing biological omics datasets, which consist of an enormous number of features with small sample sizes (Huang, 2021). Mutual information (MI) is a measure in information theory. In information theory, a correlation or relationship exists between two random variables if we can derive information about one by observing the other (Tisoc et al., 2022). Figure 3 shows a pictorial representation of MI, where A and B are correlated random variables. The intersection between individual entropy H (A) and the circle of individual entropy H (B) implies the equivalence between mutual information I (A, B) and conditional mutual information J (A, B). The joint entropy H (A, B) is the joint entropy of the random variables A and B.

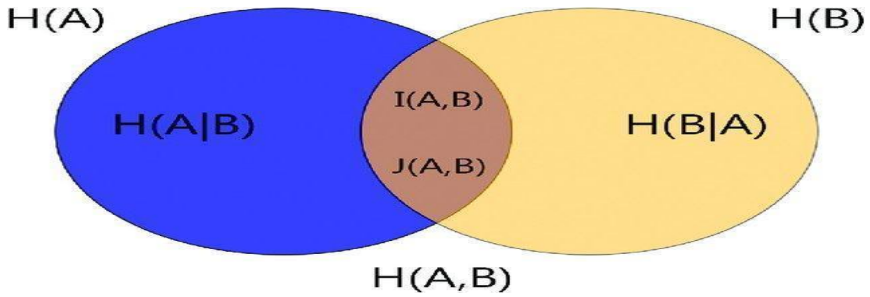

Figure 3: Schematic Representation of Mutual Information Technique (Parvandeh et al., 2020).

As stated in the data source section, a total of 4931 expressed genes is still a huge number of variables when compared to the total sample available. Besides, most of these features(genes) have little or no relevance. Thus, retaining the most informative and relevant features (genes) with the strongest correlation with the target variable remains a substantial aspect of this research. Therefore, Mutual information (MI) was applied to our gene count matrix to reduce its high

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

dimensionality, given a set of our label data. We used the "*mutual_info_classif*" method from the scikit-learn library to select the best features with higher correlation with the target data.

Nested Cross Validation and Forward Features Selection

Feature selection and classification techniques most often work in a mutually dependent manner in high-dimensional data. It is good practice to filter out irrelevant features from the ML model. Explicitly, using too many irrelevant features in the training model may yield predictions that tend to validate the data, as the bias-variance trade-off leans toward high variance (Tisoc et al., 2022).

There are several approaches to address this problem. In our study, Nested Cross-validation (nCV) in conjunction with Forward Features Selection (FFS) was applied to select the best features and to constrain the classifiers as much as possible to reduce overfitting. This technique uses a series of train, validation, and test set splits.

In the inner loop, the score is maximised by fitting a model to each training set, which is directly used to select hyperparameters on the validation set, and generalisation error is calculated by averaging test set scores across several dataset splits in the outer loop. The dataset is split into k-outer folds (3-folds), each of these folds is used for training, with K-1 split into the inner fold for training and testing in 5-inner nCV. The outer loop was used for the evaluation of our model, and the inner loop for feature selection and hyperparameter tuning. Grid search with different parameters in each of the 5 inner folds (nested) was manually implemented.

A model was trained with a K-1 split on the best hyperparameter from innerCV and tested on the test data split in the outer CV-fold of unseen data, with the average of the accuracies obtained from the 5 iterations. The entire process repeats itself for the available outer-folds. For the feature selection, in each iteration, a feature (gene) is selected with constrained tuning as well as the model accuracy. In subsequent iterations, we keep adding a variable (gene) to the already selected feature without selecting the already appended gene. This implies that for each nCV, 5 different features (genes) were selected. The figure below demonstrates the architecture of nCV FFS model used in this research.

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

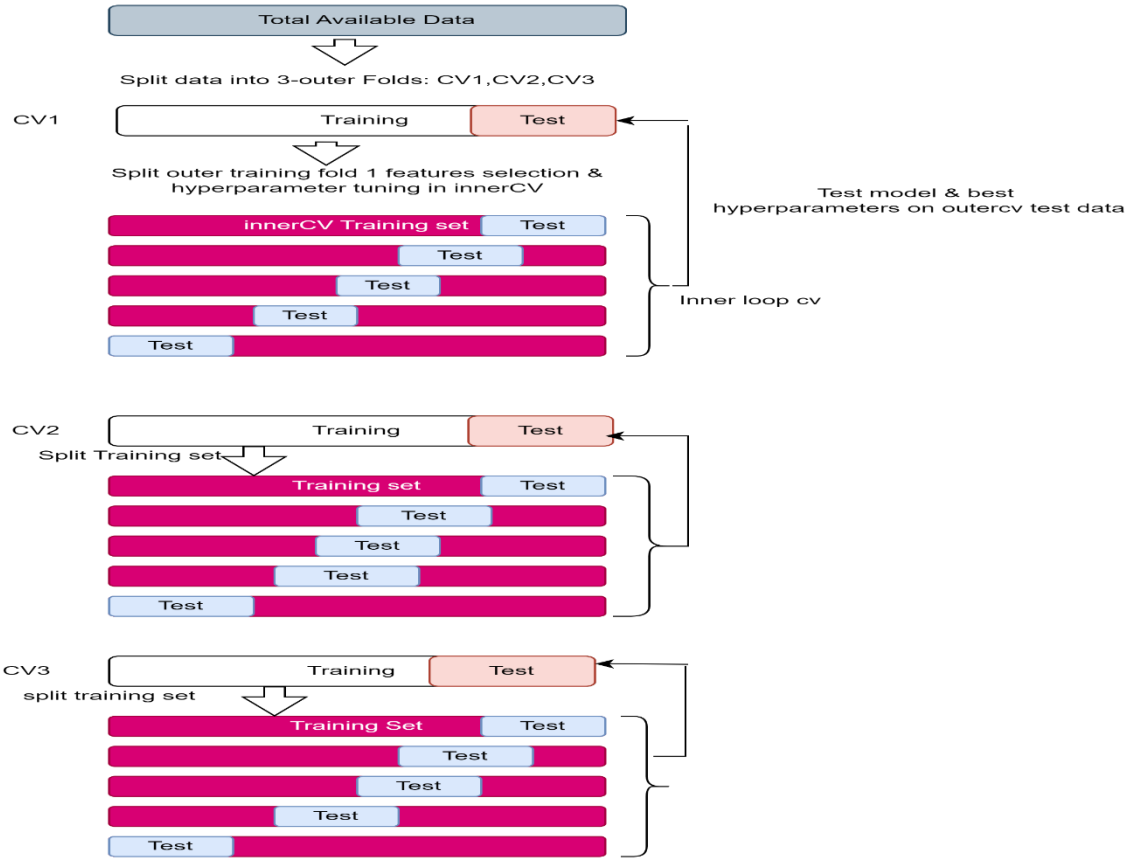

Figure 4: Nested Cross-Validation Schematic Representation

Machine Learning Classification Algorithms

This study is solely based on binary classification. The ML classification algorithms applied are RF, SVM, KNN, Logistic Regression, and GB. All the necessary libraries for executing our model were based on the Scikit-learn library in Python. As stated in the data source section, our label data had four stages (Stage I, Stage II, Stage III, and Stage IV). For better binary classification and performance, the observations of Stage I were replaced by Stage II because there are very few (i.e. 4) data available in Stage I. Thus, Stage II, III, and IV were applied to ML classification models with possible combinations as Stage II vs Stage IV, Stage II vs Stage III, and Stage III vs Stage IV by splitting the dataset with the corresponding gene count matrix dataset. For each binary classification, the above-mentioned ML models were aggregated using the nCV technique with different hyperparameter tuning.

Hyperparameters are parameters that can be constrained or adjusted and fine-tuned to improve the performance of the ML model. There are a handful of hyperparameter settings available for

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

different ML algorithms. We manually applied a grid search of different parameters for the 5 iterations in each inner fold. The same approach was adopted for other folds, but with different tuning.

RESULTS AND DISCUSSION

As stated earlier, the aim of this study is to investigate the feasibility of classifying BCa stages using RNA-Seq TPM count data from the TCGA database with ML approaches. The initial approach to this was to get an insight into the data distribution. Therefore, Figures 5, 6, and 7 demonstrate the visualisation result from the t-SNE methodology on how the target data (bladder cancer stages) are segregated from each other.

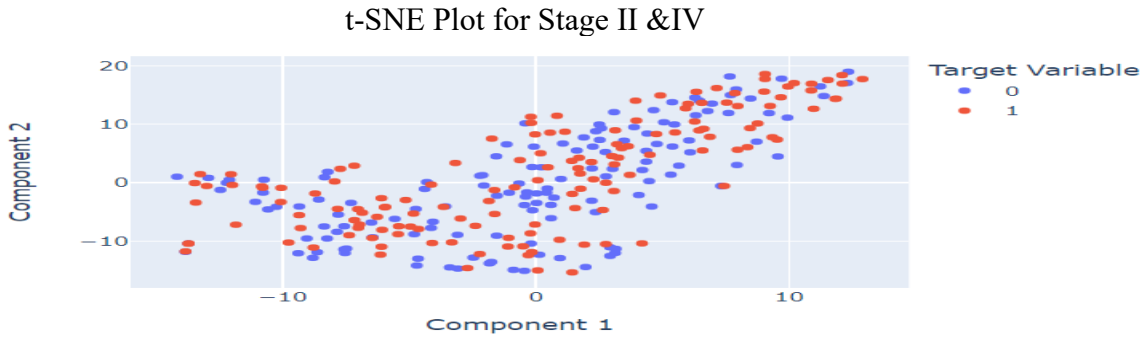

*Figure 5: t-SNE data Distribution for Stage II&IV. (**0**: Stage II, **1**: Stage IV)*

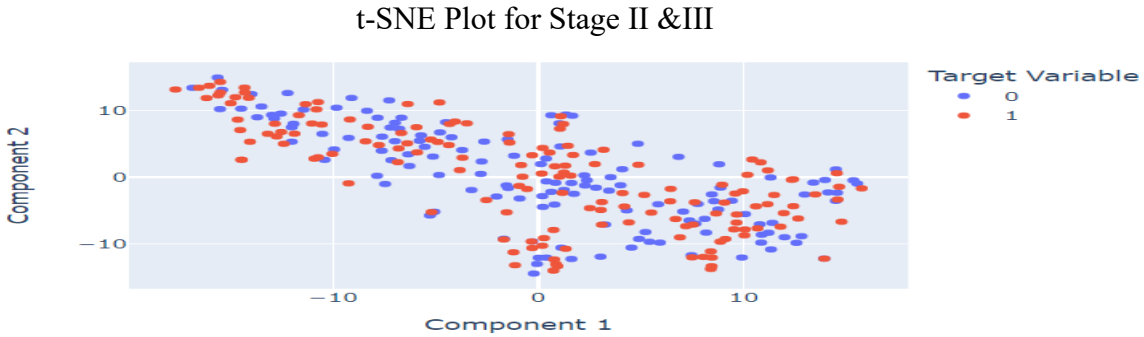

*Figure 6: t-SNE data Distribution for Stage II&III. (**0**: Stage II, **1**: Stage III)*

t-SNE Plot for Stage III &IV

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

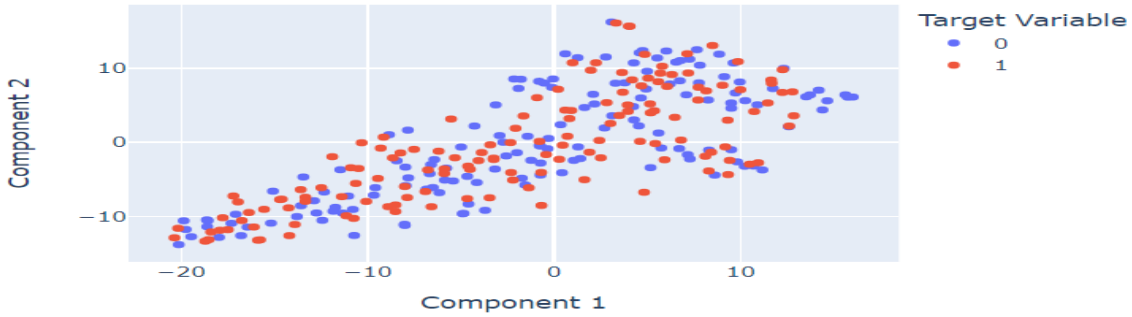

*Figure 7: t-SNE data Distribution for Stage III&IV. (**0**: Stage III, **1**: Stage IV)*

Due to the high dimensionality of our dataset, we applied mutual information for dimensionality reduction to obtain relevant features (genes) for our ML models. Hence, Table 4 presents an extract from the list of features with stronger correlations with our class labels, as determined by MI.

Table 4: List of Features (genes) Determined by Mutual Information Technique

ADIPOR2	VCL	MRPS10	PARP12	NFYC	PTGS2	TP53INP2	
CD59	TMEM40	RGS1	SORBS1	HNRNPM	RANBP1	PGK1	TLE5
HSPB1	RARRES2	CXCL12	UNC5B	ALDH3A1	BST1	ATP5F1B	H3C3
BCL6	NR4A3	CXCR4	CD244	SOX4	ERGIC3	IDO1	BARX1
RAB25	HMGCS2	FST	ARGLU1	EHF	CKAP4	DDX56	SLCO2B1
RAB15	CYP1A2	RETREG3	S100A8	RAB5A	SCUBE3	INA	PRSS2
MMP3	BATF	CLIC6	TMEM69	TLCD1	KCNF1	RPL22L1	H2AC8
C4orf3	GEM	PSIP1	HNRNPK	CCT2	GPX4	TUBA1A	FEN1
FOS	PFKFB3	RPS9	PWWPB	CYP4F11	GNG12	RNF213	IGHV1-69-2
SNHG29	SFN	ASCL3	MYLPF	NAA38	FHL3	H2AC20	SNN
KRT10	HES4	SUMO2	NCOA6	RNU5D-1	C2orf72	PSMB10	
MT-TC	MT-TT	IGKJ1	IGHV3-11	RNU6-876P	EMP2	UBA52	
IGHV4-59	NDUFAF8	MTCO2P12	RTL9	APOBECC	CWC25	EPOP	

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

The predicted features from MI techniques are attached with relevance scores based on their information scores, which range from 0 to 1. The closer it is to 1, the better the correlations between the variables. Figure 8 visualises a few of the listed features according to their MI scores.

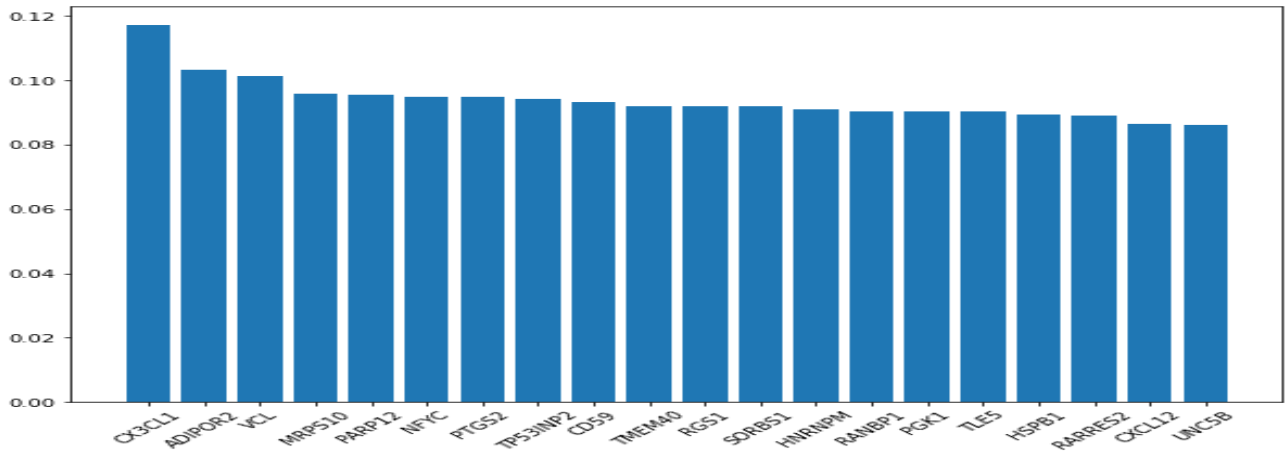

Figure 8: A Plot of MI Features Plot based on their Scores.

Model Performance Evaluation

Our ML model classification reports are presented in Table 5. The results presented are deduced from the average performance of the 3-outer-folds from the ML classifier that had better results from the three binary classifiers. It states how well our model performed based on *Precision*, *Sensitivity*, and how accurate the classification of the different BCa stages is. “0 and 1” are the label encoders assigned to each label (BCa stage).

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Table 5: ML Model Classification Performance Report

Binary Classifications BCa stages	Outer Folds	ML Model	Average Precision	Average Sensitivity	Best Accuracy
II: 0	1-3	RF	**0.51**	0.51	0.60
IV: 1			0.54	0.54	
II: 0	1-3	SVM	0.47	0.17	0.51
III: 1			0.52	0.84	
III: 0	1-3	GB	0.53	0.58	0.55
IV: 1			0.53	0.46	
II: 0	1-3	LR	0.53	0.49	0.60
III: 1			0.54	0.64	

The figures shown below were the outcome of our ML model evaluation reports of our test data. We used a confusion matrix and AUC to assess how well our model performs. We reported the models with better results, a confusion matrix, and plots of true positive rate (sensitivity) against false positive rate (specificity) for our binary classes. Figures 9, 10, 11, and 12 show ML model evaluation for BCa cancer stage II and IV with labels A, B, and C representing confusion matrix reports for outer-fold 1, 2, and 3, respectively. The ROC plots are a combination of each of the ROC plots from the three outer folds. The confusion matrix tables show an uneven distribution in *TP*, *FP*, *TN*, and *FP* predicted cases. In outer-fold 2 of RF classifiers, 19 cases of stage II were predicted as *TP*, and 28 as *FP*. While 18 cases of BCa stage IV were predicted as *FN*, and 28 as *TN*. A similar trend in results is observed in other ML performance evaluation reports.

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

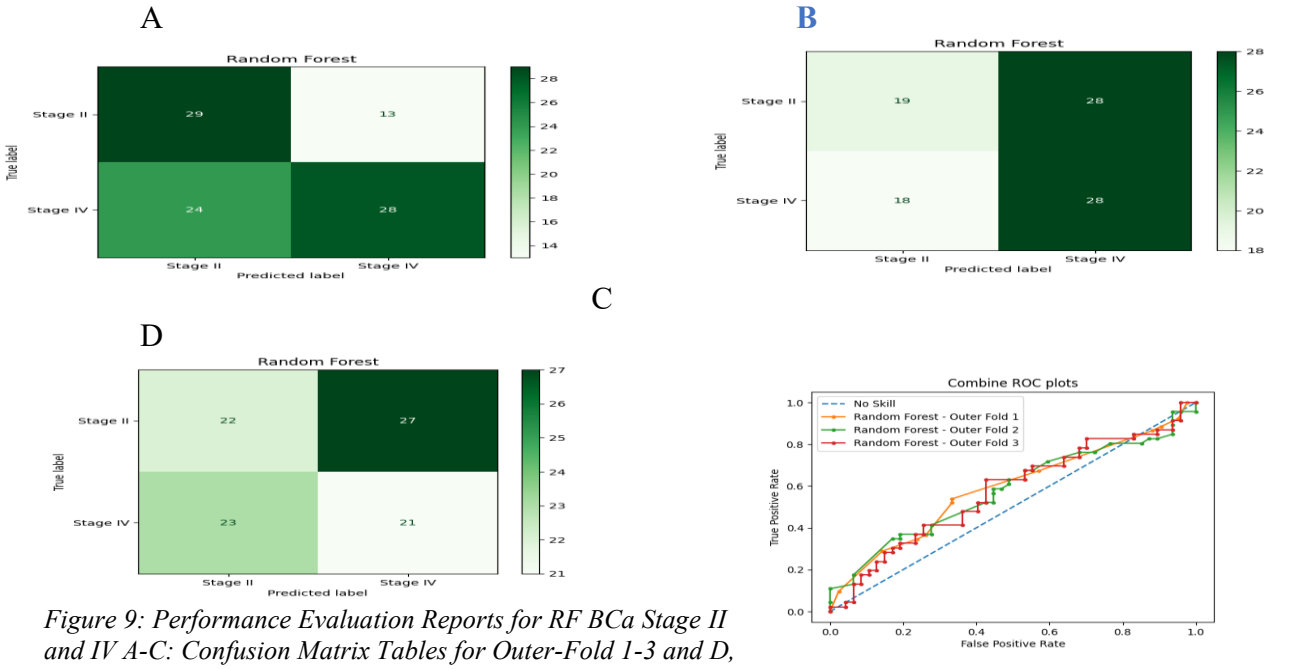

Figure 9: Performance Evaluation Reports for RF BCa Stage II and IV A-C: Confusion Matrix Tables for Outer-Fold 1-3 and D, the corresponding ROC plots

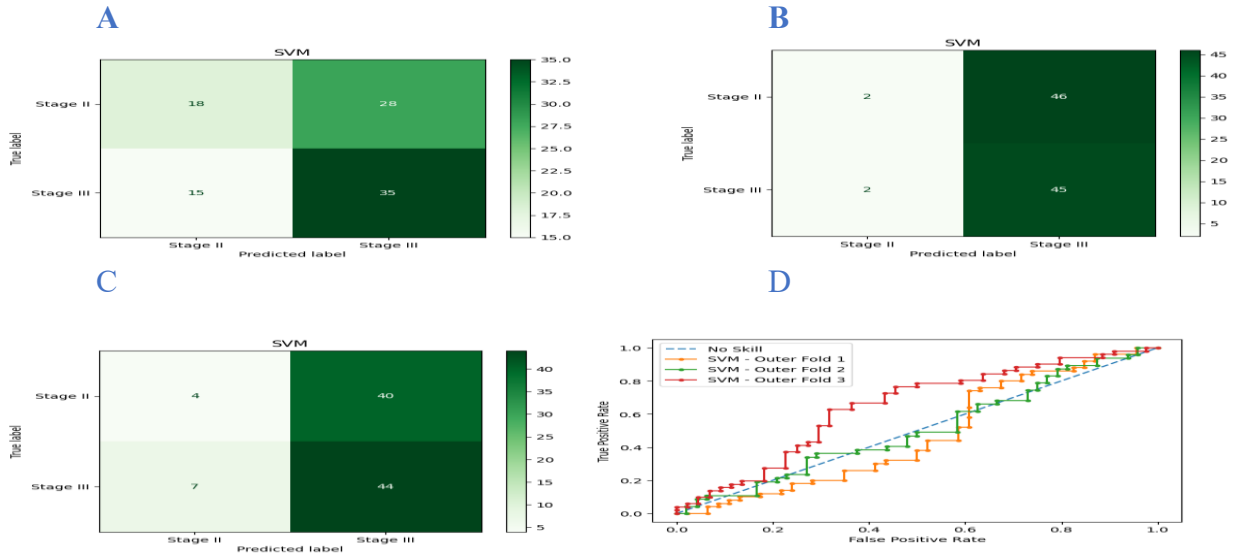

Figure 10: Performance Evaluation Reports for SVM BCa Stage II &III
A- C: Confusion Matrix Tables for Outer Fold 1-3 and D, the corresponding ROC Plots

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

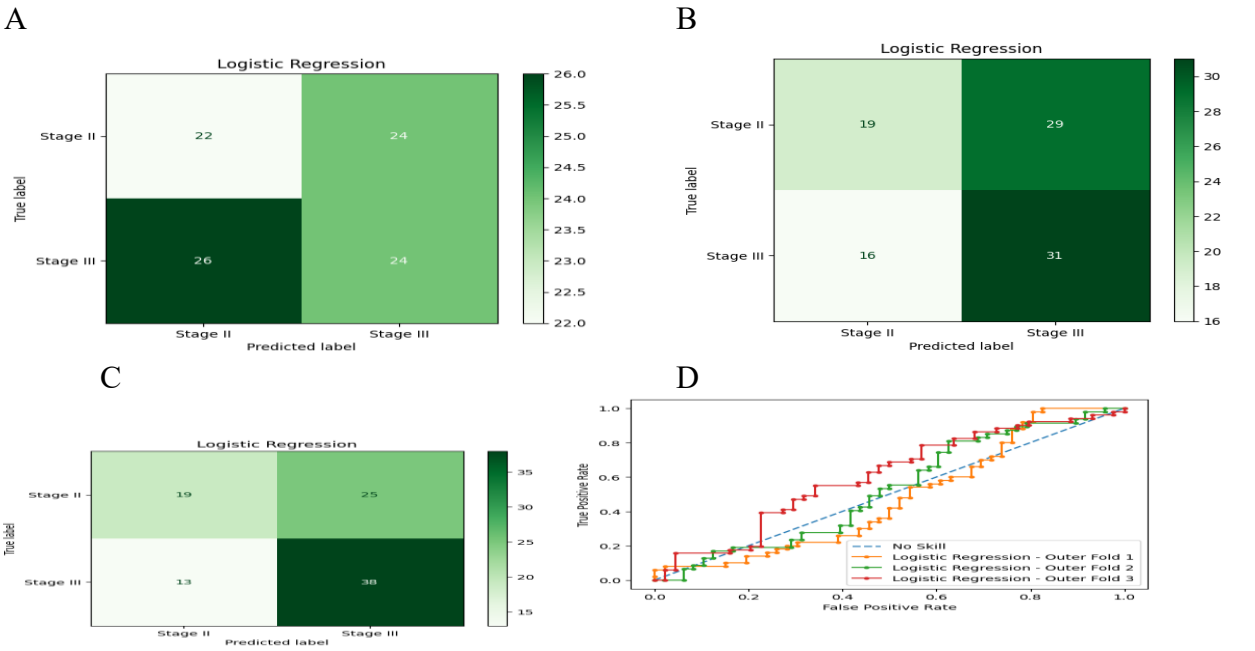

Figure 11: Performance Evaluation Reports for LR BCa Stage II &III
A-C: Confusion Matrix Tables for Outer Fold 1-3 and D, the corresponding ROC Plots

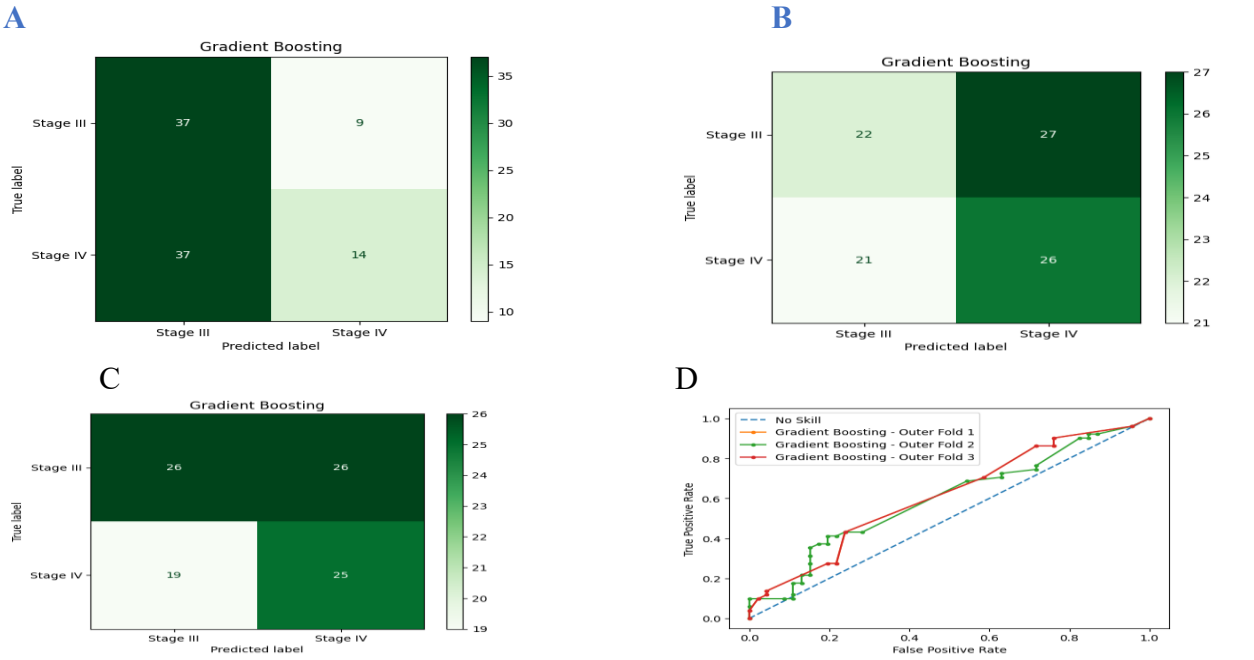

Figure 12: Performance Evaluation Reports for Gradient Boosting BCa Stage III &IV
A-C: Confusion Matrix Tables for Outer Fold 1-3 and D, the corresponding ROC Plots

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

The findings of this study add to a growing body of work suggesting that gene expression data do not align well with traditional clinical staging systems. In bladder cancer specifically, the consensus molecular classification developed by Kamoun et al. (2020) highlights that molecular subtypes provide a more meaningful way to distinguish tumours than staging alone. This supports the idea that stage progression may not be directly captured at the transcriptomic level. The relatively low predictive performance observed in this study (AUC 0.55–0.61) is therefore not entirely unexpected.

Previous work by Zhao et al. (2020, 2021) has highlighted limitations of TPM normalisation, particularly when comparing samples across datasets, which can affect downstream analyses such as classification. In addition, technical issues, such as batch effects reported in the TCGA bladder cancer data by Wang Q et al. (2018), may further blur the already subtle differences between adjacent stages, as observed in this research. The work of Zheng et al. (2025) also suggests that bulk RNA-seq data may miss important biological signals due to cellular heterogeneity, with single-cell approaches revealing patterns that are not detectable in bulk data.

These findings suggest that bulk transcriptomic data alone may have limited value for distinguishing between closely related clinical stages, consistent with the work of Zhao et al. (2021) and Zheng et al. (2025). Instead, they reinforce the view that such data are likely to be more useful for identifying molecular subtypes or predicting clinically relevant outcomes, particularly when combined with other data types or more integrative approaches.

CONCLUSION

Cancer is a life-threatening disease. As a result, attention is drawn from the medical field with more research on how to mitigate it. The field of Bioinformatics has proven useful in discovering new knowledge and insights about cancer, including what causes cancer, how to predict cancer stages/grades, and, ideally, aid in its possible treatment. Throughout this research paper, the aim is to investigate the likelihood of predicting BCa stages using ML algorithms based on RNA-Seq TPM count data and their clinical pathological stages as target data from the TCGA database.

In line with the goal, we applied 3-fold nested cross-validation and forward feature selection with five ML classification techniques (RF, SVM, GB, LR, K-NN) to our datasets. To estimate the performance of our models, a confusion matrix table and AUC from ROC curves were adopted. Our work yielded average AUCs of 0.58, 0.54, and 0.61 for stage II&IV, II&III, and III&IV, respectively, using RF, SVM, and GB. Also, 54% and 64% as the highest precision for BCa stage IV and stage III for RF and LR, respectively, were determined. Similarly, 54% and 64% *Sensitivity* for RF and LG, respectively, with an overall accuracy of 60%.

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Bulk RNA-seq transcriptomic profiles were not very effective at distinguishing adjacent bladder cancer stages, with AUCs ranging from 0.55 to 0.61. But this limited performance actually tells us something important: the boundaries between cancer stages are not defined by distinct gene expression patterns in the primary tumour. Rather, these stage transitions seem to depend more on anatomical factors, the extent of tumour spread, and changes in the tumour microenvironment. Better approaches for future studies might include predicting patient survival directly, classifying tumours by molecular subtype, integrating multiple molecular data types, or using single-cell sequencing to capture the complexity that bulk tissue analysis misses.

This study highlights the challenges of predicting clinical stages from TPM-normalised transcriptomic data and underscores the need for computational approaches that more accurately reflect the underlying biology of bladder cancer progression. In practical terms, these findings suggest that bulk RNA-seq TPM transcriptomic data should not be relied on solely for staging bladder cancer in clinical or predictive settings.

Instead, more informative approaches such as combining molecular subtypes, integrating multi-omics data, or focusing on models that predict clinical outcomes are likely to provide greater value for decision-making and future research. It was revealed that bulk RNA-seq profiles struggle to distinguish between bladder cancer stages, serving as a vital 'reality check' to the medical community.

It suggests that relying too heavily on current transcriptomic models could lead to errors in staging bladder cancer and in its treatment plan. Further research to improve precision medicine for bladder cancer should combine gene expression with clinical variables or explore molecular features beyond bulk TPM data. This integrated approach may provide more reliable predictions to guide treatment choices.

REFERENCES

Bosserhoff, A., & Kappelmann-Fenzl, M. (2021). Next generation sequencing (NGS): What can be sequenced? In M. Kappelmann-Fenzl (Ed.), Next generation sequencing and data analysis: Learning materials in biosciences. Springer. https://doi.org/10.1007/978-3-030-62490-3_1

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2. https://doi.org/10.1177/117693510600200030

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2023). Global cancer observatory: Cancer today. International Agency for Research on Cancer. https://gco.iarc.fr/

Garapati, S. S., Hadjiiski, L., Cha, K. H., Chan, H. P., Caoili, E. M., Cohan, R. H., Weizer, A., Alva, A., Paramagul, C., Wei, J., & Zhou, C. (2017). Urinary bladder cancer staging in CT

Copyright © African Journal of Applied Research

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

urography using machine learning. Medical Physics, 44(11), 5814–5823. https://doi.org/10.1002/mp.12510

Goutas, D., Tzortzis, A., Gakiopoulou, H., Vlachodimitropoulos, D., Giannopoulou, I., & Lazaris, A. C. (2021). Contemporary molecular classification of urinary bladder cancer. In Vivo, 35(1), 75–80. https://doi.org/10.21873/invivo.12234

Guo, C. C., Bondaruk, J., Yao, H., Wang, Z., Zhang, L., Lee, S., Lee, J. G., Cogdell, D., Zhang, M., Yang, G., Dadhania, V., Choi, W., Wei, P., Gao, J., Theodorescu, D., Logothetis, C., Dinney, C., Kimmel, M., Weinstein, J. N., McConkey, D. J., & Czerniak, B. (2020). Assessment of luminal and basal phenotypes in bladder cancer. Scientific Reports, 10(1), 9743. https://doi.org/10.1038/s41598-020-66747-7

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Medicine, 9, 75. https://doi.org/10.1186/s13073-017-0467-4

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., & Zhang, H. (2020). RNA sequencing: New technologies and applications in cancer research. Journal of Hematology & Oncology, 13, 166. https://doi.org/10.1186/s13045-020-01005-x
https://doi.org/10.1016/j.csbj.2014.11.005
https://doi.org/10.1093/bib/bbz081

Huang, Z. (2021). Comparison of mutual information-based feature selection method for biological omics datasets. In Proceedings of the 8th International Conference on Soft Computing & Machine Intelligence (ISCMI) (pp. 60–63). IEEE. https://doi.org/10.1109/ISCMI53840.2021.9654940

Kamoun, A., de Reynies, A., Allory, Y., Sjödahl, G., Robertson, A. G., Seiler, R., ... & Weinstein, J. (2020). A consensus molecular classification of muscle-invasive bladder cancer. European urology, 77(4), 420-433.

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nature Communications, 10(1), 5416. https://doi.org/10.1038/s41467-019-13056-x

Kong, C., Zhang, S., Lei, Q., & Wu, S. (2022). State-of-the-art advances of nanomedicine for diagnosis and treatment of bladder cancer. Biosensors, 12(10), 796. https://doi.org/10.3390/bios12100796

Kourou, K., Exarchos, K. P., Papaloukas, C., Sakaloglou, P., Exarchos, T., & Fotiadis, D. I. (2021). Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal, 19, 5546–5555. https://doi.org/10.1016/j.csbj.2021.10.006

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015).

Kumar, Y., Gupta, S., & Singla, R. (2022). A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. Archives of Computational Methods in Engineering, 29, 2043–2070. https://doi.org/10.1007/s11831-021-09648-w
Machine learning applications in cancer prognosis and prediction: A systematic review. Computational and Structural Biotechnology Journal, 13, 8–17.

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Parvandeh, S., Yeh, H. W., Paulus, M. P., & McKinney, B. A. (2020). Consensus features nested cross-validation. Bioinformatics, 36(10), 3093–3098. https://doi.org/10.1093/bioinformatics/btaa046

Qbal, M. J., Javed, Z., & Sadia, H. (2021). Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. Cancer Cell International, 21, 270. https://doi.org/10.1186/s12935-021-01981-1

Saginala, K., Barsouk, A., Aluru, J. S., Rawla, P., Padala, S. A., & Barsouk, A. (2020). Epidemiology of bladder cancer. Medical Sciences, 8(1), 15. https://doi.org/10.3390/medsci8010015

Shastry, K. A., & Sanjay, H. A. (2020). Machine learning for bioinformatics. In K. Srinivasa, G. Siddesh, & S. Manisekhar (Eds.), Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications: Algorithms for intelligent systems. Springer. https://doi.org/10.1007/978-981-15-2445-5_3

Song, H., Yang, S., Yu, B., Li, N., Huang, Y., Sun, R., Wang, B., Nie, P., Hou, F., Huang, C., Zhang, M., & Wang, H. (2023). CT-based deep learning radiomics nomogram for the prediction of pathological grade in bladder cancer: a multicenter study. Cancer imaging : the official publication of the International Cancer Imaging Society, 23(1), 89. https://doi.org/10.1186/s40644-023-00609-z

Tisoc, M., Marcelo, B., & Jhosep. (2022). Mutual information: A way to quantify correlations. Revista Brasileira de Ensino de Física, 44. https://doi.org/10.1590/1806-9126-rbef-2022-0055

Toh, C., & Brody, J. P. (2021). Applications of machine learning in healthcare. Smart Manufacturing: When Artificial Intelligence Meets the Internet of Things, 65.
TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data. Briefings in Bioinformatics, 21(6), 2223–2234.

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B. E., Iacobuzio-Donahue, C. A., Betel, D., Taylor, B. S., Gao, J., & Schultz, N. (2018). Unifying cancer and normal RNA sequencing data from different sources. Scientific data, 5, 180061. https://doi.org/10.1038/sdata.2018.61

Wang, Y., Mashock, M., Tong, Z., Mu, X., Chen, H., Zhou, X., Zhang, H., Zhao, G., Liu, B., & Li, X. (2020). Changing technologies of RNA sequencing and their applications in clinical oncology. Frontiers in Oncology, 10, 447. https://doi.org/10.3389/fonc.2020.00447

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

Wigner, P., Grębowski, R., Bijak, M., Saluk-Bijak, J., & Szemraj, J. (2021). The interplay between oxidative stress, inflammation and angiogenesis in bladder cancer development. International Journal of Molecular Sciences, 22(9), 4483. https://doi.org/10.3390/ijms22094483

ISSN: 2408-7920
Copyright © African Journal of Applied Research
Arca Academic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 170-192
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Xu, X., Xie, Z., Yang, Z., Li, D., & Xu, X. (2020). A t-SNE based classification approach to compositional microbiome data. Frontiers in Genetics, 11, 620143. https://doi.org/10.3389/fgene.2020.620143

Zhao, Y., Li, M. C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. Journal of translational medicine, 19(1), 269. https://doi.org/10.1186/s12967-021-02936-w

ISSN: 2408-7920
Copyright Ⓒ African Journal of Applied Research