Journal of Engineering Science and Technology
Vol. 19, No. 2 (2024) 725 - 743

AN EARLY RNA-SEQ DETECTION SYSTEM FOR BREAST TUMOURS BASED ON MACHINE LEARNING

A. M. ALEESA[1,*], AHMED A. MOHAMMED[2], YAHYA AHMED[3]

[1]Department of Programming, College of Computer Science and Information Technology, University of Kirkuk, 52001, Kirkuk, Kirkuk Governorate, Iraq
[2]Department of Computers and Information, College of Electronics Engineering Ninevah University, Mosul, Nineveh, 41002, Iraq
[3]Presidency Department, Northern Technical University, Mosul 41002, Iraq
*Corresponding Author: ahmed.marwan@uokirkuk.edu.iq

Abstract

Cancer, a pervasive global health issue, accounts for approximately 9 million deaths annually. The survival rate of cancer patients significantly improves with early detection and accurate staging. In this context, ribonucleic acid sequencing (RNA-Seq) has become a powerful technique for measuring gene expression, thereby playing a crucial role in human disease research. On the other hand, there is a need for more efficient computational resources and tools for analysing RNA-Seq data. The RNA-Seq datasets known as the Cancer Genome Atlas (TCGA) were used in this research. In contrast, The following five types of cancer are included: Colon Adenocarcinoma, Prostate Adenocarcinoma, Renal Clear Cell Carcinoma, Lung Adenocarcinoma, and Breast Invasive Carcinoma. This research proposes a machine-learning technique based on the AdaBoost classifier for detecting, classifying, and predicting breast cancer. The findings of our proposed method exhibit remarkable performance, achieving a cross-validation accuracy of 99.77%, while the test and prediction accuracy were 100%. Critical parameters such as precision, recall, support, F1-score, and accuracy support this performance.

Keywords: Adaboost, Breast tumours, Cancer detection, Cancer genome atlas,. Machine learning, RNA-sequence.

1. Introduction

Cancer is the second leading cause of death in many nations, following congestive heart failure disease [1]. Worldwide, the number of cancer patients has dramatically increased. According to the International Agency for Research on Cancer's latest survey, the number of new cancer cases in 2020 reached 22.4 million, with 6.2 million deaths. In countries with low and moderate incomes, cancer fatalities occur around 70%. Lack of exercise, obesity, tobacco use, vegetable intake, and soft fruit and alcohol use are five lifestyle variables that contribute to about one-third of cancer-related diseases. Cigarette smoking is the leading cause of cancer, accounting for around 43% of cancer-related deaths among smokers [2]. Cancer is characterised by uncontrolled cell growth and may occur in any body part. Cancer cells can grow and spread to other regions of the body. Cancer is the same, yet it alters and extends in different ways. It all starts when cells leave equilibrium, resulting in aberrant cell multiplication. As cells grow in the body, they obstruct normal physical activities. Cancer may affect many organs, including the colon, breast, lungs, blood, and large intestine. Some tumours have modest cell division, whereas others have rapid cell proliferation. In particular, colon cancer has emerged as one of the leading causes of cancer-related fatalities globally. However, in its early stages, this disease frequently goes unrecognised, making it difficult for patients to recognise their condition [1].

The TCGA is a significant milestone in genomics research, covering many topics, including protein expression, somatic transitions, copy number variation, gene expression, microRNA expression, and DNA methylation. Eleven thousand primary cancer patients' molecular profiles have been sequenced and described. The National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) established TCGA, a joint endeavour known as pilot research, in 2006. It focused specifically on three forms of cancer: lung, glioblastoma, and ovarian. Because of the substantial accomplishments made during the initial phase, NCI and NHGRI reauthorised TCGA for a full-scale development process in 2009. The TCGA collected data from over 11,000 cases spanning 33 tumour types over a decade, resulting in an extensive and vast dataset that tracks the molecular abnormalities in cancers [3, 4]. These massive datasets have created enormous prospects for classifying global abnormalities at the RNA, DNA, and protein levels [5].

Beyond the encouraging results that have been shown in the application of RNA-Seq in cancer diagnosis, still there are several obstacles to overcome, a significant challenge remains in the use of RNA-Seq for cancer classification. Another most challenging issue is the numerous genes in the data samples produce a high dimensionality of gene expression data [6]. Another difficult challenge might be in effective data management and analysis. Also, the findings' dependability, durability, and explainability raise additional concerns. More comprehensive and diverse datasets are urgently needed [6]. Additionally, the prolonged duration of the process remains a significant problem, due to the vast number of genes required for classifying human cells [7].

These weaknesses may impair the models' functionality and restrict their applicability to an extensive variety of populations. Better computational resources and tools are also crucial for analysing RNA-Seq data [8]. As a consequence, there is a research gap that requires an investigation for more efficient and effective approaches for cancer classification employing RNA-Seq data. due to Machine

learning capability for extracting indefinable patterns from large datasets, it is considerably aids in the early diagnosis of breast cancer. This is particularly useful in genomics, where simultaneously measuring expression levels of thousands of genes is necessary [9, 10]. Machine learning algorithms can quickly analyse vast amounts of data more efficiently than traditional methods. Correspondingly, Machine learning models improve with increasing exposure to data. As the availability of genomic data expands, these models are anticipated to achieve increasing accuracy in cancer prediction [9, 11, 12].

The AdaBoost technique has garnered considerable attention as a machine learning classification method, primarily due to its minimal error rate. This characteristic renders it particularly suitable for datasets with low noise [13, 14]. As a boosting algorithm successor, it combines a group of weak classifiers to form a classifier model that achieves more robust prognosis results [13]. The AdaBoost algorithm has been successfully utilised in various scientific trials to solve problem classes in object recognition, including facial, visual, and signal processing systems. As a result, numerous researchers have effectively employed the AdaBoost algorithm in addressing challenges related to object identification, such as face detection in images and videos.

In the present study, we employed the AdaBoost algorithm to classify publicly available data extracted from the TCGA Pan-Cancer dataset [15], segregating it into distinct tumour categories. Specifically, patient samples obtained from this comprehensive dataset were utilised. The primary objective of this investigation is to offer a broad outlook on the predictive determinants for patients diagnosed with early-stage cancer while also conducting a comparative analysis of the model's performance against precise calculations.

2. Related Work

The healthcare domain stands out as an exceptionally suitable sector for the application of data science, owing to the abundance and relevance of available data types. Within hospitals, data flow constitutes a dynamic process primarily consisting of numerical values. Healthcare, as an open framework, provides fertile ground for the development of data analysis and machine learning research. Dhar posits that proficiency in computer science equips researchers with valuable insights and the ability to make data-driven predictions [16]. Numerous researchers have made available datasets about breast cancer, most of which possess a level of accuracy conducive to classification tasks [17, 18].

Deshpande et al.[19] discussed the immense popularity of RNA-seq due to the continuous efforts of the bioinformatics community to develop accurate and scalable computational tools to analyse the enormous amounts of transcriptomic data it produces. The study also highlights the potential of RNA-seq analysis in detecting novel exons or whole transcripts, assessing the expression of genes and alternative transcripts, and studying alternative splicing structures. However, One drawback could be that it lacks practical examples and case studies to demonstrate the use of these tools [19]. Carangelo et al. [20] described how single-cell RNA sequencing (scRNA-seq) has become a prevalent and effective method for biomedical research, enabling the profiling of the complete transcriptome of numerous individual cells and highlighting the diversity of intricate clinical samples. The most impressive aspect of this study is its thorough analysis of the

scRNA-seq industry, exploring the latest developments and potential applications of these techniques. However, a potential weakness could be that it does not provide practical examples or case studies to illustrate the application of these tool.

Dindhoria et al. [8] discussed the significant innovations in next-generation sequencing techniques and bioinformatics tools that have impacted our understanding of RNA. It also explains various computational resources, tools, and bioinformatics analyses advancement for small and large non-coding RNAs. The strength of this study is its comprehensive overview of the computational approaches for non-coding RNA identification and annotation. However, a potential weakness could be that it does not discuss these tools' practical applications or implications in biological research.

Hsu and Si [21] focused on categorising 33 cancer patients using RNA sequencing findings from the Cancer Genome Atlas (TCGA). Linear support vector machine (linear SVM), decision tree (DT), Artificial neural network (ANN), k-nearest neighbours (KNN), and polynomial support vector machine (poly SVM) are the five machine learning techniques created. The experimental findings revealed that linear SVM achieved the most favourable results, demonstrating a classification accuracy of 95.8% within the research context. Lyu and Haque [22] proposed a novel approach to identifying potential biomarkers for each cancer type. Their model capitalised on a comprehensive understanding of 33 common cancer tumour categories, drawing upon the pan-cancer atlas. A visual neural network was used to recognise tumour morphologies and show the neural network's outputs, finding the genes with the most tremendous significance in tumour categorisation. Using high-dimensional RNA-Seq data, two-dimensional representations were created, allowing the categorisation of 33 distinct cancer tumour types. Using the Guided Grad Cam approach, a noticeable heat map for each gene was developed for each class. When using a train/test split, the suggested technique attained a staggering classification accuracy of 95.59%.

O et al.[23] used blood-based gene expression signatures and machine-learning techniques to build an approximation strategy for classifying transcripts. RNA data from the Omnibus database of Gene Expression was used, together with machine learning methods written in R. The study demonstrated relatively robust discrimination of autism-specific conditions through cluster analysis. A support vector machine and a k-nearest neighbour classifier were employed to validate data outcomes, yielding high accuracy rates of 93.8%, 87.5%, and 100% for accuracy, specificity, and sensitivity, respectively. Qi et al. [24] shed light on the advantages and limitations of clustering and classification methods applied to recent advancements in integrating, reporting, and retrieving single-cell RNA sequencing (scRNA-seq) data. The study encompassed linear and non-linear approaches, employing dimensionality reduction techniques tailored to scRNA-seq data.

Wenric and Shemirani [25] implemented enormous ensembles of RNA-Seq genes to construct a supervised learning approach for gene-collecting samples. Then, the random forest classification approach was used to produce arbitrary ranking measures and extracted feature rankings from 323 to 1210 cancer RNA sequencing datasets (which includes extreme pseudo-samples) using Variational Autoencoders with Regressors. The research found latent potential in supervised learning algorithms for gene selection in RNA-Seq training, highlighting the importance of gene selection strategies in genetic expression analyses.

Song et al. [26] focused their research on building a classification strategy for assessing cancer gene expression data. They used a hybrid recursive exclusion function with the Adaboost method to uncover relevant classification characteristics. The research revealed significant advancements. Tarek et al. [27] developed gene expression cancer classification data and presented an organisational group classification strategy that improved classification performance and balance. The results demonstrated that ensemble classifiers rely less on individuals from a single group.

The AdaBoost approach has recently gained attention as a hybrid method for machine learning, given its low error rate that aligns well with low-noise data collections [13, 14]. It combines a series of weak classifiers to construct a model that achieves superior prediction outcomes, representing an advancement over boosting algorithms. Consequently, numerous research studies have successfully utilised the AdaBoost algorithm in various domains, including visual, video sequences, and signal processing systems, to address classification challenges in object detection. For example, Zhou and Wei [28] employed the AdaBoost algorithm to extract the top 20 core features from the Xm2VT Face Database, revealing a 54.23% reduction in calculation time. Furthermore, Sun et al. [29] utilised the AdaBoost algorithm to extract high-order features and weights from the UCI machine-learning repository. The results demonstrated that the integrated classifiers alone outperformed the HPWR classification methods in accuracy. However, only some studies have explored the application of AdaBoost and random forests for medical database prediction.

In recent research, Gupta and Gupta [30] employed Artificial Neural Networks (ANN) as a prediction technique for mesothelioma, they had accomplished an outstanding accuracy rate of 96%. Their research focused on cancer classification using gene expression data, and they employed many techniques to enhance the classification process. They performed a logarithmic transformation on the gene expression data to preprocess it and make the classification procedure less complicated. The Bhattacharya distance metric was also used to select the study's most important genes. Gradient Descent and the Genetic Optimisation Algorithm (GOA) are used by the Deep Belief Neural Networks weight update mechanism to determine the average error. Testing on datasets related to leukaemia and colon cancer proved the effectiveness of the proposed cancer classification method. Using gene expression data, the classification method produced an impressive 0.9534 accuracy rate and 0.9666 detection rate. These findings demonstrate the effectiveness of their method for accurately classifying and identifying cancer based on patterns of gene expression.

3. Method

This section reviews the methodology used in this research. We explain the dataset that has been used and also pre-process the dataset to ensure its appropriateness for further analysis. This comprises techniques such as data normalisation, balance dataset classes and feature selection, all striving to improve the dataset's integrity and accuracy. Figure 1 enhances its readability and includes a detailed description of the text. The figure illustrates the workflow of our study, starting from the RNA-Seq datasets from TCGA, through preprocessing and balancing of the dataset using SMOTE, to the application of the AdaBoost classifier for cancer detection,

classification, and prediction. Each step in the workflow is now clearly labelled and discussed in the corresponding sections of the paper.

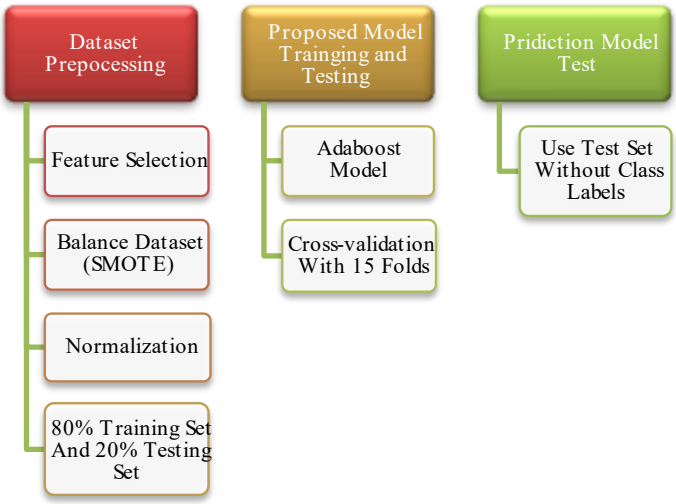

Fig. 1. System flowchart diagram.

Afterwards, we present the developed model, which was created exclusively for the categorisation of breast tumours. We discuss the underlying algorithm and its essential components, emphasising the reason for its selection and its benefits in correctly classifying tumours. Detailed insights into the model architecture, including its training and testing methods, are offered to ensure the transparency and repeatability of the experimental technique. We want to provide a firm basis for the following research and assessment of tumour classification performance by describing the breast cancer data preparation procedure and providing the created classification model. The systematic approach employed in this phase establishes the resilience and trustworthiness of our research technique, allowing for proper interpretation and assessment of experimental outcomes.

3.1. Dataset preprocessing

The preprocessing of RNA-Seq datasets is a crucial step in ensuring the accuracy and reliability of downstream analyses. This process involves several key stages:

3.1.1. RNA-seq cancer dataset

The gene expression mechanism in the RNA-Seq dataset used in this study incorporates valuable information from The Cancer Genome Atlas (TCGA) Pan-Cancer Analysts [31-33]. The dataset contains 801 samples (rows) from various types of cancer. Each instance is described by the expression levels of 20,532 genes (columns). The Irvine Machine Learning Repository provided this vast metadata, verifying its trustworthiness and accessibility for research purposes.

The data variable contains the entire set of gene expression values from all 20,532 genes. It specifically offers detailed information on the expression levels of each gene in the 801 samples. Furthermore, each of the 801 samples has associated

labels that designate particular cancer kinds. As shown in Table 1, these designations are expressed as strings corresponding to the corresponding cancer-type acronyms.

A required change is accomplished using a "Label Encoder" approach to incorporate these labels into later evaluation methods as shown in Table 1. This method guarantees that the cancer-type abbreviations have been converted into an appropriate numerical representation that can be effectively used in the dataset's evaluation and analysis.

Table 1. Dataset details with encoded label.

Abbreviations	Encoded Label	Dataset Instances
BRCA	0	300
COAD	1	78
KIRC	2	146
LUAD	3	141
PRAD	4	136

We enable detailed research and investigation of the complicated connections between gene expression patterns and cancer types by exploiting this large gene expression RNA-Seq cancer dataset and applying relevant pre-processing techniques. The particular consideration that we contribute to data quality, labelling, and encoding enhances the robustness and dependability of our future analyses, allowing for correct interpretations and meaningful conclusions on the relationship between gene expression patterns and cancer classification [34].

3.1.2. Feature selection

The feature selection technique used in this research is the feature importance method with the ExtraTree classifier used to calculate the Gini Importance for each feature. This ensemble learning technique aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its classification result It is very similar to a Random Forest Classifier and only differs from it in constructing the decision trees in the forest [35-38].

The Gini Importance, also known as the mean decrease impurity, measures the total reduction of the criterion brought by that feature. It is also known as the total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples running that node)) averaged over all trees of the ensemble [38].

After calculating the Gini Importance for each feature, features were ranked based on their importance scores. For further analysis, a threshold has been set to select features with an importance level above 0.002. The number of 108 features has been chosen based on the point of features' importance. This approach reduces the dimensionality of the dataset, improving computational efficiency and potentially enhancing model performance by eliminating irrelevant or redundant features [34, 35].

3.1.3. Balancing a dataset

Balancing a dataset is crucial in machine learning, especially when dealing with imbalanced datasets where one class significantly outnumbers the other [39]. This

imbalance can lead to a model that performs well on the majority class but poorly on the minority class. One popular technique to address this issue is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE is an oversampling method that balances class distribution by randomly increasing minority class examples by replicating them. It works by selecting samples close to the feature space, drawing a line between the samples in the feature space and drawing a new sample at a point along that line [40].

Specifically, for each instance in the minority class, the k-nearest neighbours are k-nearest neighbours computed. Then, depending on the amount of oversampling required, one or more are selected to create synthetic examples. To make all dataset labels equal to the highest label, we used SMOTE to oversample the minority classes until they contained the same number of examples as the majority class, as shown in Fig. 2. This helps to establish a balanced dataset, which may boost the classifier's performance However, it's vital to realise that SMOTE might increase the performance of models on unbalanced datasets. Still, it doesn't guarantee optimum outcomes and should be regarded as one component within a broader arsenal of strategies for handling imbalanced data [40].

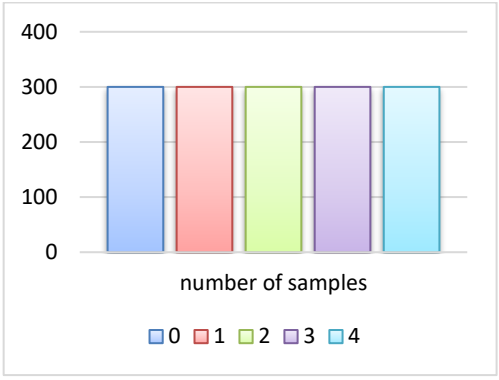

Fig. 2. Number of instances per class.

3.1.4. Normalization

Normalisation is a key step in data preprocessing, particularly when dealing with features that have different scales and units. The MinMax Scaler is a common normalisation method; it adjusts features by scaling each feature to a specified range, usually between 0 and 1. The formula gives this adjustment:

$$X_{std} = (X - X_{min(axis=0)}) / (X_{max(axis=0)} - X_{min(axis=0)}) \quad (1)$$

$$X_{scaled} = X_{std} * (max - min) + min \quad (2)$$

where min, max = feature range. This adjustment is often used instead of zero mean, unit variance scaling. Outliers do not influence the MinMax Scaler. Regardless, it linearly adjusts them down into a particular range, with the largest occurring data point being the maximum value and the smallest indicating the smallest value. It's important to note that you should fit the MinMax Scaler using the training data and then apply the scaler to the testing data before the prediction. This ensures that the test set during training doesn't influence the model, which can lead to overfitting.

3.2. Machine learning technique

The following sections will delve into the specifics of each machine learning technique used in this study, including their theoretical background, implementation details, and performance evaluation metrics. We will also discuss how these techniques are applied to our dataset and how they contribute to achieving our research objectives.

3.2.1. AdaBoost classifier

AdaBoost (Adaptive Boosting) is a machine-learning method used as a classifier—additionally, AdaBoost is a well-established machine-learning algorithm known for its simplicity in design. AdaBoost combines multiple weak classifiers to form a robust classifier. Each weak classifier is simple to understand and interpret, making the overall model easier to comprehend. When used in conjunction with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm, causing later trees to focus on mistakes made by previous trees (i.e., it adapts to the learner's weaknesses). Within the framework of the AdaBoost method, a collection of weak classifiers is carefully selected and merged to create a practical evaluation module for cancer data classification. Let

$$H = (\tilde{h}f) \tag{3}$$

represent the set of weak classifiers. The training dataset consists of feature-label pairs, denoted as $[(x_1, y_1), (x_i, y_i), (x_n, y_n)]$, where xi represents the i^{th} feature $y_i \in (+1, -1)$ indicates the label of the i^{th} feature vector, indicating whether the feature vector exhibits normal or abnormal behaviour. The value of n represents the size of the dataset. Additionally, let (w_1, w_i, w_n) represent the weights assigned to the samples, which signify their significance and serve as a statistical approximation of the sample distribution. This summarises the AdaBoost algorithm.

Step (1) Initialize weights

$$w_i = (i = 1,\ldots n) \tag{4}$$

subject to

$$\Sigma_i^n{=}1\ w_i(1) = 1 \tag{5}$$

Step (2) Iterate for $(t = 1,\ldots T)$.

(a) Compute the weighted errors ε_j for each weak classifier h_j

$$\varepsilon_j = \Sigma_{i=1}^{n} w_i\,(t)\ I\ [y_i \neq h_j(x_i)] \tag{6}$$

where $I[\gamma]$ is an indicator function defined as:

$$I[\gamma] = \begin{cases} 1, & \gamma = True \\ 0, & \gamma = False \end{cases} \tag{7}$$

Select a poor h(t) classifier, which minimises weighted classification errors from the constructed weak classifier

$$h(t) = \arg\min \varepsilon_j. \tag{8}$$

$h_j \in H$

(b) Select the weak classifier h(t) with the minimum weighted classification error $\varepsilon(t)$.

(c) Compute the weight α(t) assigned to the selected weak classifier

$$\alpha(t) = \frac{1}{2}\log\left(\frac{1-\varepsilon(t)}{\varepsilon(t)}\right) \quad (9)$$

(d) Update the weights w_i for the next iteration. Update the weights by

$$w_{i(t+1)} = \frac{w_i(t)\ exp\ (-\alpha(t) y_i h(t)(x_i))}{Z(t)} \quad (10)$$

where the normalization factor Z(t) is given by:

$$Z(t) = \sum_{k=1}^{n} w_k(t)\ exp\ (-\alpha(t) y_i h(t)(x_k)) \quad (11)$$

Step (3) The strong classifier H(x) is characterized by:

$$H(x) = \text{sign}\ (\sum_{t=1}^{T} \alpha(t) h(t)(x)) \quad (12)$$

The AdaBoost-based approach identifies cancer by combining categorical and continuous variables into a robust classifier, ensuring a natural interaction between both feature types without any artificial transitions. This is a critical factor in the algorithm's effectiveness. Notably, the weighted error classification rate of the effective classifier improves as the number of iterations (T) in the AdaBoost method grows. When T approaches infinity, the process significantly reduces the weighted error rate while the error rates of the weak classifiers stay below 50% [41, 42]. This strong convergence ensures the algorithm's dependability. Furthermore, we considerably reduce false rates related to standard samples and distinct cancer kinds by utilising decision stumps. The careful selection of weak classifiers guarantees their error rates are continually kept below 50%, leading to the algorithm's overall convergence. Rudin et al. [42] investigated the AdaBoost algorithm to grasp its convergence features fully. Furthermore, AdaBoost has fewer hyperparameters to tune, making it easier to implement than other complex models. This simplicity in implementation reduces the chances of errors and increases the efficiency of the model development process. Table 2 illustrates the critical parameters of AdaBoost.

Table 2. Adaboost parameters.

Parameter	Type
n_estimators	200
learning_rate	0.1
estimator	None
algorithm	SAMME.R

3.2.2. Model evaluation

Our machine learning models' performance is measured using numerous measures, including Accuracy, Precision, Recall, and F1 Score. These metrics give a complete picture of the model's performance across several dimensions.

Confusion Matrix: The confusion matrix is used to assess the effectiveness of the classification model. The confusion matrix is based on the number of accurate and wrong guesses, summarised using count values and split down by class [43, 44]. The requirements are as follows:

False Negative (FN): When the classifier predicts the samples are false, but actually, it's true.

False Positive (FP): When the classifier predicts the samples are true, but actually, it's false.

True Negative (TN): When the model classified the actual value, it is negative.

True Positive (TP): When the model classified the actual value as True [45].

The formula for accuracy is:

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{total\ pridiction} \quad (13)$$

- **Precision**: Precision metrics indicate how many accurately predicted instances were positive. These measures decide whether or not the model is dependable [36]. The formula for precision is:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \quad (14)$$

- **Recall (Sensitivity)**: The number of really positive instances that might be reliably predicted using the model is represented by recall metrics [36]. The formula for the recall is:

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negative\ (FN)} \quad (15)$$

- **F1 Score**: F1 provides a combined concept of accuracy and recall measures. When we attempt to increase the precision value, the recall decreases and vice versa [36]. The formula for F1 Score is:

$$F1\ Score = 2\ \frac{Precision * Recall}{Precision + Recall} \quad (16)$$

Machine learning widely uses these metrics and provides a robust measure to evaluate model performance.

3.2.3. Model validation

The Repeated Stratified K-Fold function is a powerful tool in machine learning for model validation. Repeated stratified k-fold cross-validation was performed on the training dataset of the Adaboost model with 15 folds and one hundred repeats to reduce the overfitting problem of the Adaboost model, meaning one class has many more samples than the other [46]. Here's a brief explanation of the parameters you've used:

- n_splits=15: The dataset will be split into 15 folds or subsets. In each iteration, 14 subsets will be used for training the model, and one subset will be used for testing.
- n_repeats=100: the model training will be repeated 100 times. This helps obtain a less biased model, as each data point will be in the testing set 100 times and in different folds.
- random_state=30: This is used for initialising the internal random number generator, which will decide the data splitting into folds. Providing a fixed number (like 30 in this case) ensures that the function's output remains constant across multiple calls.

This method is a great way to ensure that our model is validated thoroughly and can provide reliable performance metrics. It's beneficial when working with datasets where certain classes dominate over others.

4. Result and Discussion

Tumours are traditionally diagnosed based on anatomical and protein expression characteristics observed through histology and immunohistochemistry. However, these routine histopathological methods face challenges in detecting poorly differentiated cancers and do not reveal the underlying genetic aberrations or biological pathways involved. Although diagnostic classification methods based on gene expression signatures have been identified, a comprehensive classification system that leverages gene expression data for cancer diagnosis has been developed in this research. The genes contributing to this classification have been identified, leading to an insightful understanding of the molecular basis of cancer.

Remarkably, our developed cancer classification system employs the AdaBoost algorithm with gene expression data extracted from RNA-Seq tumour cells. Further optimisation steps need to be explored to evaluate the effectiveness and success of this proposed design.

In Fig. 3, the confusion matrix for the model's performance on an 80% training and 20% for testing and validation split is presented, a standard tool for evaluating the performance of a classification model. A visual representation of the relationship between a dataset's predicted and true labels provided by a confusion matrix. The diagonal entries in this matrix reflect the number of occurrences when the predicted label equals the real label, showing accurate predictions. In contrast, the off-diagonal components show erroneous predictions in the cases when the real and predicted labels do not match. Additionally, the diagonal values are high, indicating higher accuracy predictions. This shows that the classification model has performed well on the dataset. However, it isn't easy to provide a more detailed interpretation without the actual confusion matrix or additional context.

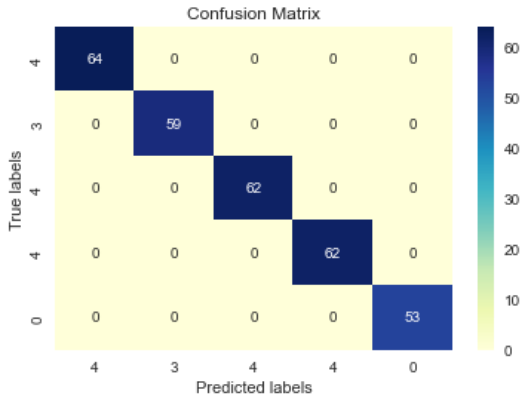

Fig. 3. Confusion matrix outcome for 80% training and 20% testing.

Table 3 comprehensively evaluates the model's performance and other metrics derived from the confusion matrix, such as precision, recall, F1-score and Support for each class in the model. Let's delve into each column. The Class column represents the distinct categories the model attempts to predict. In this case, five classes are labelled 0, 1, 2, 3, and 4. Consequently, the Precision is the ratio of correctly predicted positive observations to the total predicted positives. High

precision relates to the low false positive rate. Here, all classes have a precision score of 1.00, indicating no false positives in the predictions for any class.

Recall, also known as Sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual class. The recall score of 1.00 for all classes suggests no false negatives in the predictions for any class. Although, The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. An F1 Score of 1.00 for all classes indicates perfect precision and recall. Moreover, support is the number of actual class occurrences in the specified dataset. For instance, class 0 has a support of 64, meaning there are 64 instances of class 0 in the dataset. Accordingly, the model appears to have performed exceptionally well on the given dataset, achieving perfect precision, recall, and F1-score scores across all classes. We evaluate the efficacy of our constructed model with other models from the literature regarding classification accuracy.

Table 3. Prediction results of our proposed model.

Class	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	64
1	1.00	1.00	1.00	59
2	1.00	1.00	1.00	62
3	1.00	1.00	1.00	62
4	1.00	1.00	1.00	53

A comparative analysis of eight different machine learning models provided in Table 4, the comparison of six proposed models in various related works, accompanied by the preprocessing techniques that been used in their research with their corresponding accuracies in prediction/classification. The table starting with study conducted by Ding and Peng [47] which used the Naïve Bayes model for classification and the Minimum Redundancy Maximum Relevance (MRMR) method for feature selection to achieve an accuracy of 97.30%. Correspondingly, the second study by Mahata and Mahata [48] have achieved an accuracy of 96.77% by using a feature ranking technique as feature selection and a complement naive Bayes classifier for classification. While the third study by Bhonde et al. [49] achieved an accuracy of 97.06% by using the RNN-LSTM model as classifier and Principal Component Analysis (PCA) technique for feature selection. Moreover, the fourth study by García-Díaz et al. [50] used a Genetic Grouping Algorithm for classification and achieved an accuracy of 98.81%, while they haven't used any type of preprocessing techniques on the dataset. Yet again, in the fifth investigation [51] achieved a 98% classification accuracy by using a decremental feature selection technique as feature selection and Random Forest model for classification. Comparably, the sixth study by Salman et al.[52] achieved 96.89% accuracy in classification by using the PCA approach for feature selection and a mixture of Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM).

Furthermore, in the seventh entry our proposed AdaBoost model, which achieved a 97.81% accuracy rate without the use of feature selection, Synthetic Minority Over-sampling Technique (SMOTE), or k-fold cross-validation. lastly, an additional enhancement to our proposed model is included in the eighth component in the table, the enhancement includes balancing the dataset using SMOTE,

selecting features with high feature importance, and then using k-fold cross-validation with 15-folds, this model attained a 100% test accuracy. In addition, the significant increase in test accuracy that was demonstrated after the training stage highlights the effectiveness of our approach, with 100% prediction accuracy and 99.77% cross-validation accuracy which is the highest accuracy been achieved. It is crucial to highlight that the proposed model contains recent advances in pre-processing methods, optimisation techniques, and machine learning framework design led to significantly enhanced predictive accuracy.

Table 4. Evaluation of the developed model.

Proposed Models	Feature selection	Balancing dataset	Used ML technique	Acc.%
1. Ding et al. (2005) [47]	MRMR	NA	Naïve Bayes	97.30
2. Mahata and Mahata (2007) [48]	Feature ranking	NA	Complement Naive Bayes Classifier	96.77
3. Bhonde et al. (2022) [49]	PCA	NA	RNN-LSTM	97.06
4. García-Díaz et al. (2020) [50]	NA	NA	Genetic Grouping Algorithm	98.81
5. Venkataramana et al. (2020) [51]	Decremental feature selection algorithm	NA	Random forest	98
6. Salman et al. (2018) [52]	PCA	NA	(CNN)+(Bi-LSTM)	96.89
7. Our proposed model **without** (SMOTE, feature selection, and k-fold cross-validation)	NA	NA	Adaboost	97.81
8. Our proposed model **with** (SMOTE, feature selection, and k-fold cross-validation)	feature importance method	SMOTE	Adaboost+15 fold-cross	100

In this research, we have built a model that indicates excellent accuracy and addresses crucial practical aspects for real-world use in clinical frameworks. Initially, our developed technique is intended to be cost-effective by leverages generally accessible data, avoiding the need for extra, expensive tests or procedures. Our methodology might lower the expenses associated with misdiagnosis or delayed diagnosis by enhancing diagnostic accuracy. Additionally, the model has been designed with ease of implementation in mind and can be integrated into existing diagnostic pipelines without requiring any specialised equipment or software. Finally, our model is designed to complement current diagnostic methods, not replace them, making it an additional tool to enhance accuracy and ease integration into clinical repetition. Consequently, one of the limitations of our research is that variations in sample size can affect the robustness of our model. A smaller sample size might limit the diversity of the data and potentially introduce bias. However, we acknowledge that this might still be a limitation, and future work could benefit from even larger and more diverse datasets.

Moreover, (TCGA) is a valuable resource, but it is known to have certain biases, such as the overrepresentation of specific population groups. Consequently, it is essential to note that these biases could potentially impact the generalizability of our model. So, we have used SMOTE, and the representation for all population groups becomes the same. Also, we used a different dataset to validate the generalizability of our model. Specifically, we used a diabetes prediction dataset. Using a separate dataset is crucial in machine learning models to ensure that the model does not just fit a specific dataset's quirks but also learns general patterns that can be applied to unseen data. The validation accuracy achieved on the diabetes prediction dataset was 96.72%, indicating that our model could accurately predict diabetes status for most of the instances in the validation set; this high validation accuracy suggests that our model generalises well to new data. Furthermore, the test accuracy of our model was 96.85%. The test accuracy measures the model's performance on a separate data set not seen during training or validation. Despite its simplicity, AdaBoost is robust and performs comparably or even better than more complex models. Its ability to combine multiple weak classifiers allows it to capture intricate patterns in the data, contributing to its strong performance. These aspects make our approach practical and accessible, particularly for practitioners without advanced machine-learning knowledge. The developed technique based on the AdaBoost algorithm and gene expression data shows encouraging results in correctly classifying cancer kinds. The complete investigation of gene expression profiles gives vital insights into the molecular basis of cancer, perhaps leading to new treatment techniques.

5. Conclusion

Cancer is still a major global health problem which is responsible for approximately 9 million deaths yearly, which is uncontrolled cell development and the potential for metastasis. In cancer research, the performance and accuracy improved with the introduction of RNA-Seq which has transformed genome analysis. This research proposed an exceptional approach for categorising five forms of cancer.

The findings show that our proposed approach in categorising these various tumour kinds was attained a remarkable overall accuracy of 100%. In addition, our solution exceeds previously developed methods compared to similar studies on the accuracy of the five tumour classifications dataset. In a larger setting of cancer research our developed model can potentially contribute significantly and adjust treatment practices. This might lead to fresh insights into cancer's fundamental processes, perhaps opening up new research options. In addition, doctors may be able to begin therapy at an earlier stage of the illness in order to improve the condition of patients if an accurate and fast diagnoses been provided. An additional technologies like genetics and proteomics may give us a more comprehensive knowledge of cancer if been combined with our model. This might enable the development of medications that are adapted to the genetic composition and clinical features of each patient, boosting efficacy while reducing side effects.

On the other hand, Deep neural networks are a promising avenue for future cancer classification research and development. These networks can potentially increase accuracy and provide more complicated genomic data processing. Although our proposed approach is just one piece of the jigsaw in the battle against cancer, we feel it has the potential to make a substantial contribution. By pushing

the limits of what machine learning can do in cancer research, we want to go one step closer to a future when cancer can be successfully controlled, if not cured.

References

1. Salmi N.; and Rustam Z. (2019). Naïve Bayes classifier models for predicting the colon cancer. *IOP Conference Series: Materials Science and Engineering*, 546, 052068.
2. Gültekin, M.; Ramirez, P.; Broutet, N.; and Hutubessy, R. (2020). World Health Organization call for action to eliminate cervical cancer globally. *International Journal of Gynecologic Cancer*, 30(4), 426-427.
3. Hutter C.; and Zenklusen, J.C. (2018). The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2), 283-285.
4. Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W.K.; Luna, A.; La, K.C.; Dimitriadoy S.; Liu, D.L.; Kantheti, H.S.; and Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2), 321-337.
5. Li, Y.; Kang, K.; Krahn, J.M.; Croutwater, N.; Lee, K.; Umbach, D.M.; and Li, L. (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genomics*, 18, 508.
6. Alharbi, F.; and Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2), 173.
7. Haznedar, B.; and Simsek, N.Y. (2022). A comparative study on classification methods for renal cell and lung cancers using RNA-Seq data. *IEEE Access*, 10, 105412-105420.
8. Dindhoria, K.; Monga, I.; and Thind, A.S. (2022). Computational approaches and challenges for identification and annotation of non-coding RNAs using RNA-Seq. *Functional & Integrative Genomics*, 22(6), 1105-1112.
9. Nemade, V.; Pathak, S.; and Dubey, A.K. (2022). A systematic literature review of breast cancer diagnosis using machine intelligence techniques. *Archives of Computational Methods in Engineering*, 29(6), 4401-4430.
10. Hasan, Z.; Xing, H.J.; and Magray, M.I. (2022). Big data machine learning using Apache Spark Mllib. *Mesopotamian Journal of Big Data*, 2022, 1-11.
11. Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1), 89-109.
12. Baker, M.R.; Padmaja, D.L.; Puviarasi, R.; Mann, S.; Panduro-Ramirez, J.; Tiwari, M.; and Samori, I.A. (2022). Implementing critical machine learning (ML) approaches for generating robust discriminative neuroimaging representations using structural equation model (SEM). *Computational and Mathematical Methods in Medicine*, 2022, 1-12.
13. Ma, Y.; and Ding, X. (2003). Robust real-time face detection based on cost-sensitive AdaBoost method. *Proceedings of the 2003 International Conference on Multimedia and Expo ICME'03 Proceedings (Cat No 03TH8698)*. Baltimore, MD, USA, II-465.
14. Vezhnevets, A.; and Vezhnevets, V. (2005). Modest AdaBoost-teaching AdaBoost to generalize better. *Graphicon*, 12(5), 987-997.
15. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J.M. (2013). The

cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10), 1113-1120.

16. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64-73.
17. Bazazeh, D.; and Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *Proceedings of the 2016 5th international conference on electronic devices, systems and applications* (*ICEDSA*), Ras Al Khaimah, United Arab Emirates, 1-4.
18. Aalaei, S.; Shahraki, H.; Rowhanimanesh, A.; and Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iranian Journal of Basic Medical Sciences*, 19(5), 476.
19. Deshpande, D.; Chhugani, K.; Chang, Y.; Karlsberg, A.; Loeffler, C.; Zhang, J.; Muszyńska, A.; Munteanu, V.; Yang, H.; Rotman, J.;Tao, L.; Balliu, B.; Tseng, E.; Eskin, E.; Zhao, F.; Mohammadi, P.; Łabaj, P.; and Mangul, S. (2023). RNA-seq data science: from raw data to effective interpretation. *Frontiers in Genetics*, 14, 997383.
20. Carangelo, G.; Magi, A.; and Semeraro, R. (2022). From multitude to singularity: An up-to-date overview of scRNA-seq data generation and analysis. *Frontiers in Genetics*, 13, 2816.
21. Hsu, Y.-H.; and Si, D. (2018). Cancer type prediction and classification based on RNA-sequencing data. *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), Honolulu, HI, USA, 5374-5377.
22. Lyu, B.; and Haque, A. (2018). Deep learning based tumor type classification using gene expression data. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Washington DC USA, 89-96.
23. Oh, D.H.; Kim, I.B.; Kim, S.H.; and Ahn, D.H. (2017). Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Clinical Psychopharmacology and Neuroscience*, 15(1), 47-52.
24. Qi, R.; Ma, A.; Ma, Q.; and Zou, Q. (2020). Clustering and classification methods for single-cell RNA-sequencing data. *Briefings in Bioinformatics*, 21(4), 1196-1208.
25. Wenric, S.; and Shemirani, R. (2018). Using supervised learning methods for gene selection in RNA-Seq case-control studies. *Frontiers in Genetics*, 9, 297.
26. Song, N.; Wang, K.; Xu, M.; Xie, X.; Chen, G.; and Wang, Y. (2016). Design and analysis of ensemble classifier for gene expression data of cancer. *Advancements in Genetic Engineering*, 5(1), 1-7.
27. Tarek, S.; Abd Elwahab, R.; and Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, 18(3), 151-159.
28. Zhou, M.; and Wei, H. (2006). Face verification using gaborwavelets and adaboost. *Proceedings of the* 18th *International Conference on Pattern Recognition* (*ICPR'06*), Hong Kong, China, 404-407.
29. Sun, Y.; Wan, Y.; and Wong, A.K.C. (2006). Boosting an associative classifier. *IEEE Transactions on Knowledge and Data Engineering*, 18(7), 988-992.

30. Gupta, S.; and Gupta, M.K. (2023). Computational model for prediction of malignant mesothelioma diagnosis. *The Computer Journal*, 66(1), 86-100.
31. Fiorini, S. (2013). gene expression cancer RNA-Seq Data Set. UCI Machine Learning Repository
32. Rahman, M.; Jackson, L.K.; Johnson, W.E.; Li, D.Y.; Bild, A.H.; and Piccolo, S.R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics*, 31(22), 3666-3672.
33. Fiorini, S. (2016). gene expression cancer RNA-Seq. UCI Machine Learning Repository
34. Mohammed, A.B.; Fourati, L.C.; and Fakhrudeen, A.M. (2023). Comprehensive systematic review of intelligent approaches in UAV-based intrusion detection, blockchain, and network security. *Computer Networks*, 239,110140.
35. Sanmorino, A.; Marnisah, L.; and Sunardi, H. (2023). Feature selection using extra trees classifier for research productivity framework in Indonesia. *Proceeding of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics: ICEBEHI* 2022, Surabaya, Indonesia. 13-21.
36. Saarela, M.; and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3, 272.
37. Kharwar, A.R.; and Thakor, D.V. (2022). An ensemble approach for feature selection and classification in intrusion detection using extra-tree algorithm. *International Journal of Information Security and Privacy (IJISP)*, 16(1), 1-21.
38. Sharma, D.; Kumar R.; and Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. *Measurement: Sensors*, 24, 100560.
39. Pan, T.; Zhao, J.; Wu, W.; and Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Information Sciences*, 512, 1214-1233.
40. Pears, R.; Finlay, J.; and Connor, A.M. (2014). Synthetic minority over-sampling technique (SMOTE) for predicting software build outcomes. *arXiv preprint arXiv:*14072330.
41. Freund, Y.; Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.
42. Rudin, C.; Daubechies, I.; and Schapire, R.E. (2004). The dynamics of AdaBoost: cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5(10), 1557-1595.
43. Tasnim, A.; Saiduzzaman, M.; Rahman, M.A.; Akhter, J.; and Rahaman A.S.M.M. (2022). Performance evaluation of multiple classifiers for predicting fake news. *Journal of Computer and Communications*, 10(9), 1-21.
44. Brownlee, J. (2020). How to calculate precision, recall, and F-measure for imbalanced classification. Machine Learning Mastery. Retrieved October 5, 2023, from https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

45. Muhammad, T.; and Ghafory H. (2022). SQL injection attack detection using machine learning algorithm. *Mesopotamian Journal of Cybersecurity*, 2022, 5-17.
46. Padimi, V.; Telu, V.S.; and Ningombam, D.D. (2022). Performance analysis and comparison of various machine learning algorithms for early stroke prediction. *ETRI Journal*, 45(6), 1007-1021.
47. Ding, C.; and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2), 185-205.
48. Mahata, P.; and Mahata, K. (2007). Selecting differentially expressed genes using minimum probability of classification error. *Journal of Biomedical Informatics*, 40(6), 775-786.
49. Bhonde, S.B.; Wagh, S.K.; and Prasad, J.R. (2022). Identification of cancer types from gene expressions using learning techniques. *Computer Methods in Biomechanics and Biomedical Engineering*, 26(16), 1951-1965.
50. García-Díaz, P.; Sanchez-Berriel, I.; Martínez-Rojas, J.A.; and Diez-Pascual, A.M. (2020). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics*, 112(2), 1916-1925.
51. Venkataramana, L.; Jacob, S.G.; Saraswathi, S.; and Prasad D.V.V. (2020). Identification of common and dissimilar biomarkers for different cancer types from gene expressions of RNA-sequencing data. *Gene Reports*, 19, 100654.
52. Salman, I.; Ucan, O.N.; Bayat, O.; and Shaker, K. (2018). Impact of metaheuristic iteration on artificial neural network structure in medical data. *Processes*, 6(5), 57.