Received: October 3, 2024. Revised: November 25, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhancing Classification of Primary and Recurrent Epithelial Ovarian Cancer

Asha Abraham[1]* R. Kayalvizhi[1] Habeeb Shaik Mohideen[2]

[1]*Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India*
[2]*Department of Genetic Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India*
* Corresponding author's Email: aa6687@srmist.edu.in

Abstract: Epithelial Ovarian Carcinoma (EOC) is one of the fatal cancers, with intricate molecular features that impact its diagnosis and treatment within the female reproductive system worldwide. EOC remains one of the most lethal gynaecological malignancies, primarily due to the molecular heterogeneity in its occurrence, which complicates the classification of Primary Tumours (PT) and recurrent Tumours (RT) using traditional methods. The critical challenge lies in the size of numerous genes present in the high-dimensional data sets that possibly reduce the learning algorithm's ability and the dataset's imbalanced nature, particularly with the small number of RT samples, which led to initial models failing to classify recurrent tumours accurately. Hence, the present research introduces a Balanced EOC using the Artificial Neural Network (BEOC- ANN) model to address the challenge of accurately classifying PT and RT in EOC using RNA-sequencing (RNA-Seq) data. The dataset sourced from The Cancer Genome Atlas (TCGA) database includes 374 PT samples and 5 RT samples. A pre-processing stage is implemented, using the DESeq normalization method to handle the raw HTSeq count data and filtering 112 ovarian cancer-related genes from an initial 27,620 gene features. To resolve this, the ANN model has been fine-tuned by adding dropout layers and class weights, which helped balance the dataset. The ANN model is trained with ReLU activations for input and hidden layers, sigmoid for output, and an Adam optimizer and binary cross-entropy for loss function. A significant improvement was observed when the test size increased from 10% to 30%, allowing three RT records to be recognized. The research results demonstrate that the model achieves a training accuracy of 98%, a testing accuracy of 96%, and a recall rate of 33% for RT samples.

Keywords: Artificial neural network, Class weight balance, Epithelial ovarian cancer, Primary tumours, Recurrent tumours, Gene expression data.

1. Introduction

One challenging aspect of treating ovarian cancer (OC), the most dangerous gynaecological cancer, is the high recurrence rate that patients with this disease experience. Consequently, target therapy techniques can be improved by thoroughly understanding the genetic [1] and molecular causes of OC recurrence. Among gynaecological malignancies, epithelial ovarian carcinoma ranks second in terms of mortality. The diagnosis is advanced in about 75% of patients, making treatment challenging [2]. Given its high mortality rate and late diagnosis, ovarian cancer research is essential for early detection and treatment. Advances in surgical approaches aim to improve patient outcomes in earlier research in addition to the classification of EOC tumours in terms of primary and recurrent [3]. Cancers will be efficiently and effectively classified by utilizing microarray gene expression profiles. Due to the enormous number of genes and the minor amount of trials in gene expression information, this is a highly computational task [4]. Existing genetic testing techniques are inadequate, leaving many individuals at risk and significantly hampers prevention efforts, thus leading to inaccurate classification and missed opportunities for early intervention and improved outcomes [5].

Previous models were limited with small input samples and the lack of ability to understand the

contextual information led to misclassification. Thus, an effective Machine Learning (ML) approach employing a neural network model to classify accurately is crucial [6]. The vital aspect of maintaining ML model accuracy in classification is reducing overfitting by applying filtering methods since it performs well with high dimensional gene data [7]. These genes provide significant biomarkers for cancer diagnosis, and identification of differentially expressed genes obtained from TCGA leads to the accurate classification of Primary Tumour (PT) and Recurrent Tumour (RT) [8].

As an alternative option for existing microarrays with transcriptome analysis, the RNA-Seq method is currently being applied. It provides the accuracy required for complicated analysis, such as differential gene expression, by recording transcription activity and genome-wide gene expression. Its versatility makes it an essential tool for efficiently analyzing molecular profiles in different types of study [9]. Currently, the RNA-Seq [10] technique is faster, more accurate, more reliable, and widely used for gene expression examination to report the development of transcriptome methods. The analysis of the TCGA portal and gene expression data was retrieved with the DESeq analysis to recognize the differentially expressed genes and indicate the high-risk factors for improving survival rates of cancer patients by classifying both gene types [11]. The DESeq2 normalization method, which assumes that most genes remain unaltered, along with basic normalization techniques counts per million (CPM) and overall counts, functioned reasonably well [12]. Training a neural network for classification is performed with a class weight balancing since the minority RT samples provided more weight than the majority PT classes [13]. The dropout technique is applied to a fully connected neural network layer to estimate the prediction of machine learning model uncertainty [14]. Gene expression analysis applied for classifying cancers using ML methods helps to handle the high dimensionality of data and provides valuable data for computationally efficient analysis [15]. The main challenge in ovarian cancer classification is the imbalance in class distributions, especially the difficulty in detecting RT within a highly imbalanced dataset. RTs are considered to be less common but it is crucial for patient diagnosis, and are often underrepresented in clinical data. Traditional methods like NB and KNN techniques have struggled with accurately identifying minority classes, leads to low recall rates for RT cases. Additionally, these methods often fail to capture the complex, high-dimensional nature of gene expression data.

The innovation of this research lies in the BEOC-ANN model development, which combines an enhanced DESeq data with a focused set of filtered genes. The use of class weighting and dropout to handle the severe class imbalance achieves over 96% test accuracy and 33% RT recall. With an improved classification performance across many metrics like accuracy, better precision, recall, and f1-score when compared with the other existing models. This resultant model enhances the interpretability while maintaining high performance in distinguishing between PT and RT classes, as shown in ROC analysis.

The critical area of research contributions is given as follows:

- To enhance the classification accuracy of PT and RT in EOC using RNA sequencing data.
- To implement the DESeq normalization method to handle raw HTSeq count data from RNA sequencing.
- To address the challenge of imbalanced datasets in EOC classification that can effectively handle high-dimensional gene expression data while maintaining good generalization
- To develop a Balanced EOC using an Artificial Neural Network (BEOC-ANN) model to address the challenge of accurately classifying PT and RT in EOC.
- Dropout layers and class weights should be applied to balance the dataset and improve model performance.
- To achieve training accuracy of 98%, testing accuracy of 96%, and a recall rate of 33% for RT samples, demonstrating improved classification of RT classes.

The research article is arranged in the following sequence: section 2 follows the existing research on OC classification using ML models. Section 3 provides the research idea of the BEOC-ANN model design for classifying the PT and RT samples with the class weight balance technique. section 4 examines the results and discusses the implemented model with training and testing accuracies and Recall evaluation. Section 5 settles the research work and provides the future research scope.

2. Literature review

The RNA-Seq gene expression measures provide natural heterogeneity and noise recognition called dropouts while performing sequencing reads, hence for predicting cancer diagnosis. Kim et al. [16] applied ML methods such as Naïve Bayes (NB) and k-Nearest Neighbour implemented on binary

classification of cancer samples. Gene associations were taken out from the high-dimensional gene expression information samples obtained from TCGA databases using stacked autoencoders. A model for ovarian tumour classification utilizing radiomics and the Dual-View Global Representation (DVGR) and Local Cross Transformer (LCT) is projected by Rong et al. [17]. Deep learning with 3D CT image processing combination results in an AUC-ROC of 91.35% and an AUC-PR of 90.20%. They found that combining modern imaging techniques and machine learning enhanced cancer classification.

A Deep Semi-Supervised Generative Learning with Deep Convolutional Neural Network model (DSSGL-DCN2) is proposed by Nagarajan et al. [18] to tackle the problem of ovarian cancer classification from CT images. Experimental results demonstrate that the combined model outperforms separate networks and helps to improve cancer detection by applying deep learning to overcome obstacles like data limits and enhance classification accuracy. For malignancies with an uncertain primary origin of majority classes, Zhao et al. [19] suggested that the Cancers of Unknown Primary CUP-AI-Dx classifier uses RNA gene expression data and a 1D Inception convolutional neural network (1D-CNN) to determine the central tissue of gene. The model's performance declined to 86.96% and 72.46% on distinct datasets, indicating difficulties in real-world generalization despite achieving a top-1 cross-validated accuracy of 98.54% and 96.70% in test datasets during classification.

Laios et al. [20] compared different classifiers and feature selection strategies to apply ML for analyzing the classification performance of the 2-year diagnosis in advanced-level high-grade serous ovarian carcinoma (HGS-OC). The ensemble subspace discriminant and support vector machine techniques outperformed logistic regression, leading to an average classification accuracy of 73% with low dimensionality reduction in the feature selection process. Although the study demonstrates better prognosis accuracy, the dataset size and scope of the algorithms studied are limitations. Yan et al. [21] applied Macro ANN to simulate the immunogenomic analysis of transcriptome profiling and heterogeneity with improved outcomes of acquired RNA-Seq data from TCGA. The model also helps analyze the classification of high-dimensional data with subtype classification of OC samples. The applied prognostic factor using an ML model called MacroANN with feed-forward neural networks provides a high response rate with better clinical outcomes.

Conventional ovarian cancer classification models face huge challenges. For instance, in [16] NB and KNN techniques struggles hard with the high-dimensionality of RNA-Seq data, leading to challenges in classification accuracy and complexity in computation process. Radiomics assisted DVGR [17] model requires significant processing techniques, making them less suitable for classifying the gene expression data. The DSSGL-DCN2 method [18] offered effective cancer detection, faces challenges in handling small labeled datasets, which leads to generalization problem. The CUP-AI-Dx classifier, designed for cancers of unknown PT, lacks flexibility in classifying data from diverse gene origins [19]. SVM and ensemble classifiers relies on feature selection, omits filtering, and leads to overfitting [20]. In comparison to these, the proposed BEOC-ANN model implements gene filtering, dynamic class weighting, and advanced regularization techniques to address these above-mentioned issues, ensures accurate classification even with limited input samples and improves generalizability over conventional models. By combining gene filtering, dynamic class weighting, and advanced regularization strategies, the proposed BEOC-ANN model overcomes these obstacles. The performance of predicting ovarian cancer prognosis using RNA-Seq data, even with small sample sizes, and classification has also increased.

3. Research methodology

3.1 Data collection

The clinical data is taken from the TCGA repository (https://portal.gdc.cancer.gov/) from the Genomic Data Commons portal mentioned in [22]. Then, the information is divided into a training portion (70%) and a testing portion (30%), in which 265 samples are used for training and 114 samples for testing.

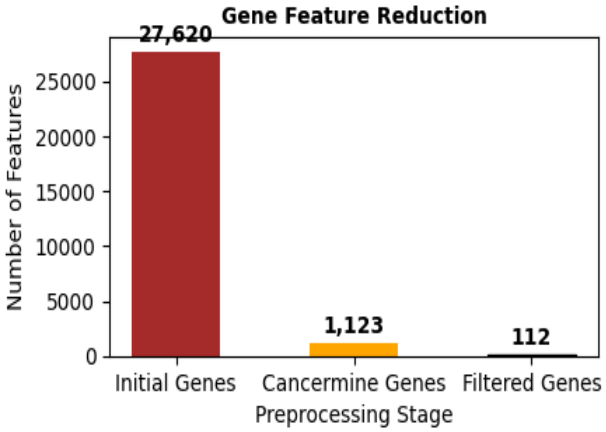


Figure. 1 Gene Feature Reduction: Pre-processing Stages

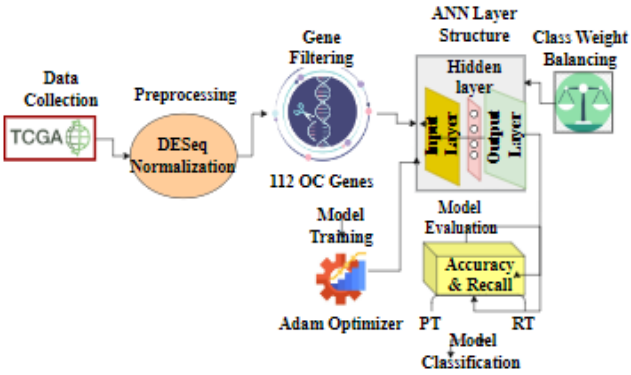


Figure. 2 Proposed BEOC-ANN Algorithm

This split ensures the model is trained on diverse samples and tested on an independent set to assess generalization. Fig.1 illustrates the reduction in gene features across three key pre-processing stages in the data analysis. The initial stage starts with 27,621 features. A collection of drivers, oncogenes, and tumour suppressor genes in cancer called CancerMine was gathered from existing research papers, including 1123 genes related to OC [23]. The filtered genes stage further refines 27,621 to 112 genes compared with the CancerMine mentioned 1123 genes, which are highlighted to emphasize their significance in the analysis.

Fig. 2 illustrates the BEOC-ANN design, which is intended to classify primary and recurring cancers in an imbalanced dataset. After collecting TCGA data, DESeq normalization is performed to eliminate inconsistencies or bias in the RNA-Seq data. Gene filtering minimizes dimensionality, concentrating on 112 OC genes.

The ANN model has a layer of inputs for these genes, a hidden layer with ReLU activation function, with an applied dropout rate of 0.3, followed by applying a class weight balancing and a binary classification output layer. Recurrent tumours are a minority class addressed by class weight balancing. Additionally, the model is skilled using the Adam optimizer and assessed for model evaluation in terms of accuracy and recall, especially the minority class called RT identification.

3.2 Data pre-processing

Initially, HTSeq counts represent the raw counts of reads like RNA sequences that map to each gene; HTSeq is a tool that provides these counts by aligning sequencing reads to a reference genome. The DESeq algorithm adjusts these raw counts to normalize the raw HTSeq count data for sequencing depth and other systematic biases, ensuring comparability across gene samples. The average geometric mean of a gene throughout all samples is used to divide the numbers for that gene in each sample. An estimated size scaling factor is calculated by taking the median of the gene ratios in a given sample and applying it to the total count of mapped reads in that sample.

$$K_{ij} = \frac{k_{ij}}{s_j \times f_i} \quad (1)$$

Where K_{ij} in Eq. (1) is the normalized count for gene i in sample j, the raw count is indicated as , k_{ij} and s_j is the size factor for sample j used for correcting the differences in sequencing depth. The variable f_i represents a gene-specific normalization factor accounting for variability in gene expression levels. The model can perform accurate downstream analysis by normalizing the gene counts using DESeq, such as identifying differentially expressed genes between samples. The ensemble IDs are unique identifiers used by the ensemble database in the form of (ENSG00000228037.1) to refer to specific genes, and these IDs are often used in raw sequencing data. The ensemble ID to common gene names, making the data more accessible to interpret and analyse. Filter out gene rows with unannotated sequences, and fewer counts, which refer to genes with shallow expression levels across the samples. These low-count filters reduce noise and focus on genes more likely to be biologically significant.

3.3 Designing an ANN model architecture

The detailed steps above explain how RNA-Seq data is collected, pre-processed and fed into a specified ANN architecture explicitly designed for classifying PT and RT in EOC. The preprocessing ensures the data is clean, normalized, and focused on the most relevant genes. The BEOC-ANN architecture has ReLU, sigmoid activations, and dropout layers. An Adam optimizer trained explicitly to handle the high-dimensional, imbalanced nature of the dataset, ultimately leading to improved classification performance. The training parameters involve ten epochs and 15 as batch size, which means it updates its parameters after processing 15 samples. This mini-batch approach helps train the model efficiently with limited memory.

3.4 Input layer

The input layer consists of 64 neurons, corresponding to the 112 filtered genes from the pre-processing stage. The activation function called Rectified Linear Unit (ReLU) calculated in Eq. (2) introduces non-linearity into the gene expression data, enabling it to learn complex patterns and lessen the

vanishing gradient problem. Gradually reduce the number of filters to 60 to 30 to capture more features and progress performance.

$$f(x) = max\,(0, x) \tag{2}$$

3.5 Hidden layer

The hidden layer consists of 32 neurons. As with the input layer, ReLU is used here to introduce non-linearity, which is crucial for distinguishing between PT and RT samples, given the data's high-dimensional and potentially non-linear nature. For regularization, the dropout technique with a rate of 0.3, 30% of the neurons in this layer are randomly set to 0 during each training epoch. The model begins with a low dropout value of 0.3 to observe baseline performance. Additionally, more dropout values, such as 0.5 and 0.7, are needed to find a balance where the model generalizes without losing too much learning capacity.

3.6 Class weight balance

The Class weights, if not adjusted, the prediction model can be biased toward the majority class and may not perform well on the minority class like RT. The experiment with weights applies different class weights, such as 70 and 74.8, and observes the changes in recall for the minority class and overall accuracy. Balancing the contribution of each majority and minority class helps the model to achieve better prediction performance and learn better from minority classes by achieving class weight balance sets, specifically at 74.8.

$$R_{mty} = N_{PT}/N_{RT} \tag{3}$$

β in Eq. (3) is a hyperparameter controlling the sensitivity of the dynamic scaling factor to the class imbalance. The parameter N_{RT} defines the no. of recurrent tumour samples and N_{PT} defines the no. of primary tumour samples. The term N_{total} indicates that the combined samples from N_{RT} and N_{PT} are given as in Eq. (4).

$$CW = \{0 :1,\ 1: \left(R_{mty} - \beta\right)\ \} \tag{4}$$

Table 1. BEOC-ANN Model Architecture

Layer Type	No. of Neurons	Activation Function	Regularization/Dropout
Input	112	ReLU	None
Hidden	32	ReLU	Dropout (rate varies)
Output	1	Sigmoid	Class Weight Balancing

The class weight for RT gives higher weight to the RT class by scaling the total sample size and normalizing it based on the majority class count N_{PT}. The dynamic scaling factor β emphasizes this weighting based on empirical results from training. This results in prioritizing the minority class influence in the learning process while maintaining a balance based on the overall class distribution and dynamic adjustment during training.

Table 1 provides a comprehensive overview of the BEOC-ANN model for classifying ovarian cancer. The input layer model comprises 112 neurons, one for each gene. Then, to avoid overfitting, 32 neurons in a hidden layer service the ReLU activation method. The output layer practices sigmoid to classify tumours as Primary Tumours (PT) or Recurrent Tumours (RT). Finally, to make the model dense, class weight balancing is employed to account for the imbalanced dataset.

3.7 Output layer

The output layer has one neuron since it is a binary classification that can be PT or RT. It uses a sigmoid activation function for binary classification, producing an expected value between 0 and 1. The classification tasks allow the model to provide a probability score indicating whether a sample is more likely to be an RT. The output yˆ which indicates the probability that the sample is of class RT is computed using Eq. (5).

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \tag{5}$$

Where $\hat{y}$ is the predicted probability that a sample belongs to the RT class, z indicates the output of the final layer that indicates the linear combination of the weights W. The term $\sigma(z)$ represents the sigmoid function that squashes the output z into a range between 0 and 1, which is interpreted as a probability the proposed model predicts the greatest possibility for being RT if the result of $\hat{y}$ is closer to 1. The proposed model predicts the greatest possibility for being PT if the value of $\hat{y}$ the result is closer to 0. A probability score $\hat{y}$ where $0 \leq \hat{y} \leq 1$, indicating the possibility that a sample is RT. Input x represents the features of the gene samples, and bias b can be calculated using Eq. (6) as follows:

$$z = W \cdot x + b \tag{6}$$

Here, W indicates the weight vector for the connections from the previous layer to the output neuron.

Pseudocode 1: *BEOC-ANN Model*

Input: *filtered genes, epochs, batch_size*

Output: *test_acc, RT_recall*

Step 1: def pre-process_data(raw_data):
normalized_data = DESeq_normalization(raw_data)
filtered_data=filter_OC_genes (normalized_data, 112)
return split_data
Step 2: def create_model(input_size=112):
model = Sequential ([
Dense (64, activation='ReLU', input_shape= (112)), Dropout (0.3),
Dense (32, activation='ReLU'), Dropout (0.3),
Dense (1, activation='sigmoid')])
Step 3: calculate $R_{mty} = N_{PT}/N_{RT}$
return (0 :1, 1: $(R_{mty} - \beta)$ *)*
Step 4: model. compile (Adam (), $\mathcal{L}$*, ['accuracy'])*
for *each epoch in range (E)* ***do***
model. fit (X_train, y_train, epochs, batch_size, validation_split, class_weight)
model. evaluate (X_test, y_test)
model. predict (X_test)
calculate test_acc, RT_recall.
end for*.*
return test_acc, RT_recall

The mathematical idea behind the pseudocode of proposed BEOC-ANN to predict cancerous gene involves with data preprocessing of genes which includes normalization and gene selection. The BEOC-ANN neural architecture implements ReLU activations in the hidden layers and a sigmoid function in the output layer, which produces a probability value between 0 and 1. The use of class imbalance handling adjusts the class weights based on the ratio of PT and RT samples to address class imbalance. The model is trained with the binary cross entropy loss function and evaluated using accuracy and RT recall metrics to assess how well it identifies samples.

The BEOC-ANN model's hyperparameters are listed in Table 2, as well as their descriptions, possible values, and ranges. For example, it specifies the number of epochs as 10, batch size as 15, and dropout rate as 0.3, all of which regulate different parts of model validation and training the BEOC-ANN model. The table also includes information about the learning rate at 0.001, which is essential for controlling overfitting, ensuring equal distribution of classes, and optimizing the model. It details the network's architecture, comprising its hidden layer of one and 32 neurons per layer and its setup and tuning parameters.

Table 2. BEOC-ANN Hyperparameter Settings

Hyperparameter	Value Used	Potential Range	Description
Number of Epochs	10	5-50	No. of complete epochs used for training.
Batch Size	15	8-32	No. of instances applied before the model update
Dropout Rate	0.3	0.2-0.7	Probability of dropping units to prevent overfitting.
Learning Rate	0.001 (default)	0.0001-0.01	Step size used in the Adam optimizer to update model weights.
Hidden Layers	1	1-3	No. of hidden layers in the network.
Neurons in Hidden Layer	32	16-128	No. of neurons in each hidden layer.

3.8 Optimization using Adaptive Momentum

Adam is an optimization algorithm chosen for its adaptive learning rates and potential for handling sparse gradients. It adjusts the learning rate for each input parameter. Adam uses AdaGrad for sparse gradients and RMSProp for online and gene expression data across samples. As in Eq. (7), it exhibits different learning behaviours.

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} \tag{7}$$

The core update rule of Adam, where the model's parameters θ are updated based on the computed gradients m_t and variance v_t, with an added term ϵ to prevent division by zero. The term θ_t represents the current weights of the model at step t. The parameter θ_{t+1} indicates the weights and biases of the BEOC-ANN that are adjusted after each iteration based on the gradients. The learning rate α with the value of 0.001 is applied, controlling the size of steps to update the weights. A dynamic and adaptive learning rate helps the model converge faster, avoiding large oscillations during training with the imbalanced gene expression data.

3.9 Loss function

Binary cross-entropy is used to evaluate how well the model-predicted probabilities match the accurate binary labels PT or RT to classify the cancer types accurately.

Table 3. List of Notations

Notation	Description	Notation	Description
K_{ij}	normalized count for gene i in sample j	θ_t	current weights of the model at step t
k_{ij}	raw count of gene samples	θ_{t+1}	weights and biases of the BEOC-ANN that are adjusted after each iteration
s_j	size factor for sample j	m_t	gradients
f_i	gene-specific normalization factor accounting for variability in gene expression levels	v_t	variance
N_{RT}	no. of recurrent tumour samples	ϵ	Added term to prevent division by zero
N_{PT}	no. of primary tumour samples	$\mathcal{L}$	loss function
N_{total}	combined samples from N_{RT} and N_{PT}	y	accurate label where 0 for PT and 1 for RT
β	dynamic scaling factor	p	predicted probability of the sample being RT
$\hat{y}$	predicted probability that a sample belongs to the RT class	TP	true positives represent the no. of RT samples accurately classified as RT
z	output of the final layer	FP	false negatives indicate the no. of RT samples mistakenly classified as PT
W	weight vector	$acc(\%)$	accuracy
$\sigma(z)$	sigmoid function	TN	true negatives that correctly forecast PT samples
x	Input	FN	false negatives in which RT samples are incorrectly predicted as PT

The loss calculates the actual label and the predicted probability variation. Minimizing this loss function during training helps the model accurate classifications.

$$\mathcal{L} = -[y * log\ (p) + (1 - y) * log(1 - p)] \quad (8)$$

where $\mathcal{L}$ in Eq. (8) represents the loss function, and y indicates the accurate label where 0 for PT and 1 for RT, p represents the predicted probability of the sample being RT.

The RNA-Seq gene expression data-based BEOC-ANN algorithm employs a neural network architecture PT and RT classification. It includes 112 input neurons, 32 neurons in the hidden layer, and a sigmoid output function. To avoid overfitting, the BEOC-ANN employs dropout and to compensate for the disparity between the PT and RT data, it uses dynamic class weighting. The model shows promise in RT case detection when trained with Adam optimization and tested with binary cross-entropy loss. Although the model is scalable and deals with class imbalance, it has limitations, such as a short dataset and an overemphasis on 112 genes.

4. Results and discussion

Through the experimental phase, various combinations of dropout rates at values of 0.3, 0.5, and 0.7 with the regularization values 0.001 and 0.01, as well as class weights, are tested to optimize the model's performance. Learning curves are monitored during training to detect signs of overfitting or underfitting, and hyperparameter tuning is performed. This analysis is crucial for evaluating the effectiveness of classification tasks.

Fig.3 illustrates the learning curves for a BEOC-ANN model trained over ten epochs, with training

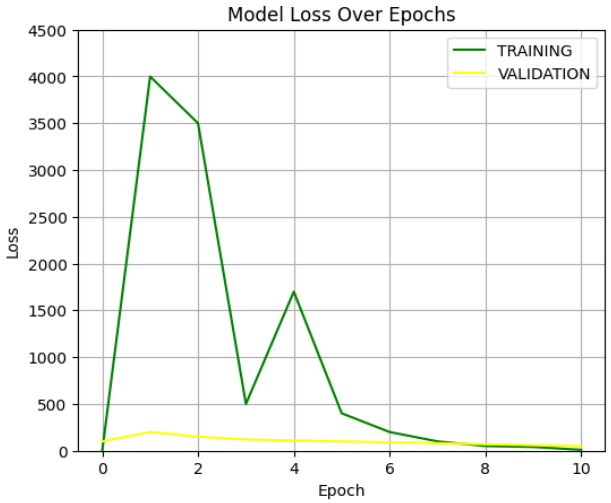


Figure. 3 Learning Curves for Loss Vs. Epochs

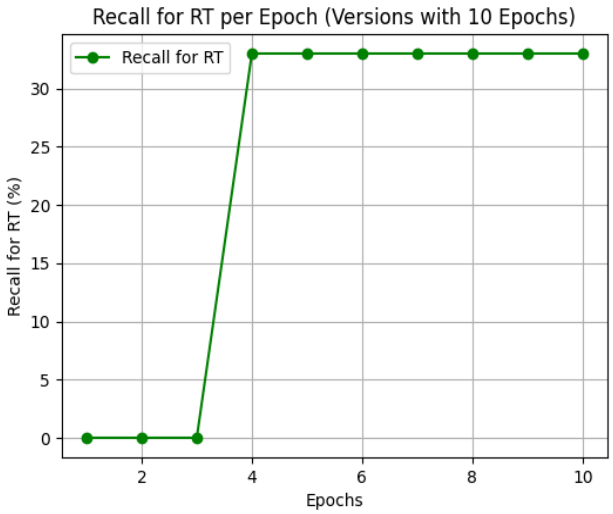


Figure. 4 Recall for RT with 10 Epochs

and validation aiding in the analysis of model convergence loss and generalization capabilities.

For this binary classification model, recall accurately analyses the actual minority classes RT samples correctly identified by the model. Recall focuses on identifying the positive class RT derived using an Eq. (9). It is instrumental in class imbalances where identifying all positive instances is critical.

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

Where TP, true positives represent the no. of RT samples accurately classified as RT. FP false negatives indicate the no. of RT samples mistakenly classified as PT.

The accuracy of correct classification occurs at 1 and incorrect as 0. The $acc(\%)$ is the ratio of correctly classified instances, PT and RT, across total samples derived using Eq. (10).

$$acc(\%) = \frac{TP+TN}{TP+FP+FN+TN} \tag{10}$$

Where TP specifies true positives correctly predicted RT samples. TN indicates the true negatives that correctly forecast PT samples. FP specifies false positives in which PT samples are incorrectly predicted as RT. FN indicates false negatives in which RT samples are incorrectly predicted as PT.

Fig.4 shows that the RT class's recall over ten training epochs is depicted in the graph with the label of Recall for RT per Epoch (Versions with 10 Epochs). The model's accuracy in identifying valid minority classes called RT cases is displayed in the graph. The X-axis represents epoch ranges with a maximum count of 10 epochs, while the Y-axis represents recall percentages for RT. The training results indicate that the low or fluctuating results may indicate problems with training, and similarly stable

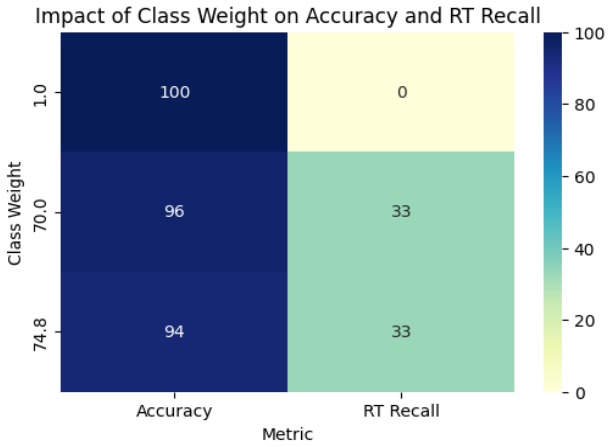


Figure. 5 Impact of Class Weight on Model Performance

or increasing recall shows successful learning of RT classes.

Fig.5 shows how different class weights affect model performance measures like accuracy and RT recall in the heatmap. The heatmap provides an easy-to-understand look at the effects of various class weights on these measures by transforming the performance data into a matrix and then visualizing it with the seaborn library. While adjusting the model's parameters, this graphic aids in determining the best class weights to strike a balance between accuracy and recall. It better helps understand and handle imbalanced datasets by adjusting class weights to increase model efficacy.

Fig. 6 explores the effect of modifying class weights on model performance. Adjusting the class weight for RT makes the model more sensitive to the minority class, improving recall. Still, it indicates potential overfitting to the minority class, suggesting a need to modify the weights to balance recall and

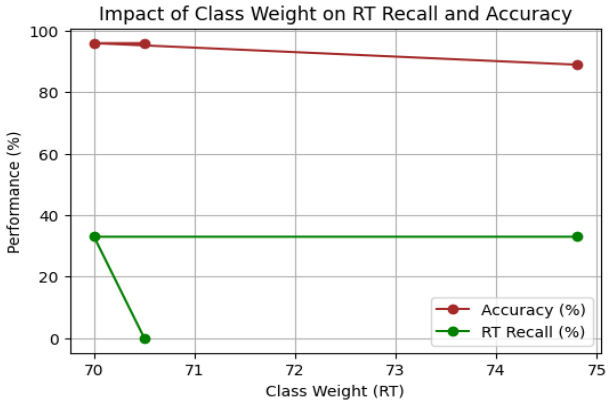


Figure. 6 Class weight on RT Recall and accuracy

Table 4. Model Versions performance comparisons

Model Version	Test Size	Drop out	Class Weight	Training Accuracy
Initial	10%	-	-	100%
Intermediate	30%	0.5	-	100%
Final (BEOC-ANN)	30%	0.3	70	98%

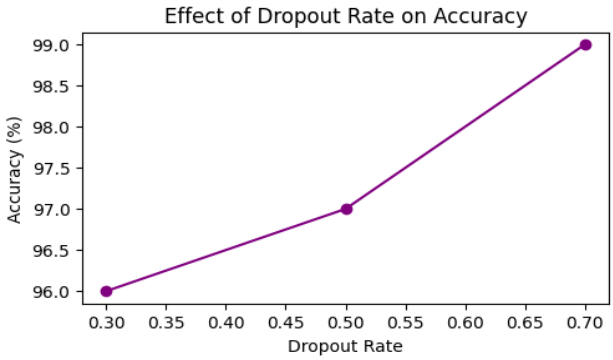


Figure. 7 Dropout effect on Model Accuracy

accuracy. As class weight is fine-tuned, recall for RT improves, reflecting the model's increased sensitivity to the minority class called RT depending on dropout adjustment too.

Table 4 provides an overview based on the model's initial, intermediate, and final (BEOC-ANN) stages. With a 10% sample size and no parameter modifications to drop out or class weights, the first model achieved 100% accuracy in the testing and training phase. Still, it failed to recognize minority class RT with a recall rate of 0%. The intermediate model kept the training accuracy at 100%. In comparison, testing accuracy dropped to 97%, and RT recall remained at 0% after increasing the test size to 30% and introducing a 0.5 dropout rate.

Fig. 7 shows the impact of the dropout rate on the accuracy of an ANN in the ovarian cancer classification task. The dropout rate plot shows that a

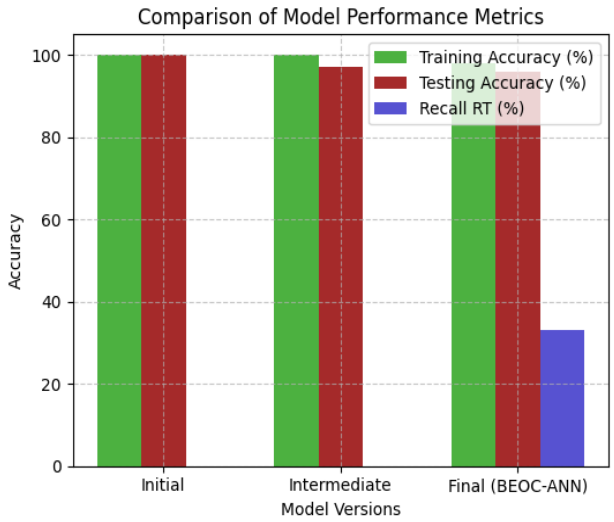


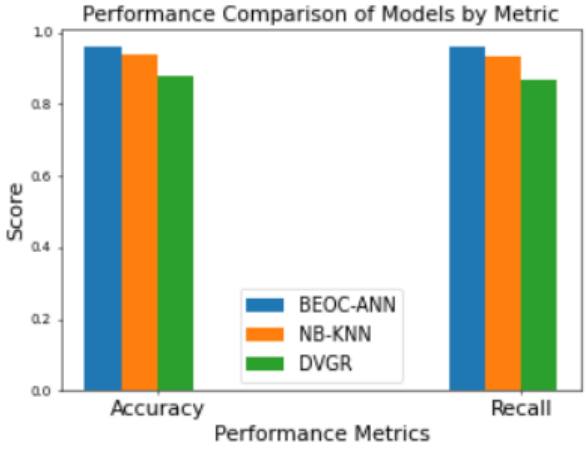
Figure. 8 Accuracy for Model Versions

moderate rate of 0.5 achieves the highest accuracy, 97%. In contrast, a higher rate at 0.7 slightly decreases accuracy, indicating that excessive neuron deactivation can hinder model performance.

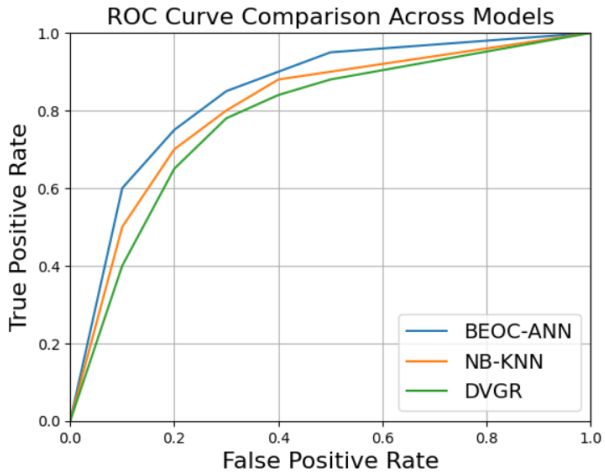
With the same test size, a reduced dropout of 0.3, and a class weight setting of 70, the Final (BEOC-ANN) model proposed an improved RT recall to 33%, training accuracy to 98%, and testing accuracy to 96%. This phase of the model result shows a balance between dropout and class weights. This final model shows how to increase RT detection while improving accuracy and EOC's overall detection capabilities.

Fig. 8 compares the model effectiveness based on an initial, intermediate, and final proposed BEOC-ANN for RT, considered a minority class, by graphing training accuracy and testing accuracy alongside recall. Overfitting is shown in the Initial model's inability to recognize RT situations despite its flawless accuracy. Although there is a slight decrease in accuracy, the Intermediate model can avoid overfitting; nonetheless, it does not have RT recall. The final BEOC-ANN model reduces dropout and improves RT recall to 33% while achieving 96% testing accuracy, thanks to class weight balancing. The system's enhanced sensitivity in recognizing recurrent cancers, which is vital for clinical applications, is reflected in this.

The three models like the proposed BEOC-ANN, NB-KNN [16] and DVGR [17] are compared in this analysis making use of a bar chart and a Receiver Operating Characteristic (ROC) curve as given in Fig. 9. The bar chart compares the model's side by side based on their scores in the metrics: accuracy and recall. The scores are scaled between 0 and 1.0. By



(a)



(b)

Figure. 9 Performance Comparative Analysis: (a) Score and (b) ROC-Curve Analysis

comparing the True Positive Rate (TPR) with the False Positive Rate (FPR) at different thresholds, the ROC curve provides a visual representation of the model's performance; a closer fit to the top left corner indicates greater classification capabilities. This comparative analysis taken as a whole, provide a thorough evaluation of the capabilities of each model.

5. Conclusion and future scope

The proposed BEOC-ANN model in this research demonstrates significant improvements in classifying primary and recurrent EOC using RNA sequencing data. The model achieved high accuracy in the training and testing phases by implementing DESeq normalization, gene filtering, and various ML techniques such as dropout, and class weight balancing. Notably, the model improved the recall rate for recurrent tumours to 33%, addressing the challenge of the imbalanced dataset in cancer classification. The study highlights the significance of ANN in handling complex, high-dimensional gene expression data for cancer classification. The improved detection of RT cases, critical for patient prognosis and treatment planning, represents a significant advancement.

Future research could improve RT classification accuracy by increasing data through synthetic techniques like the Synthetic Minority Over-sampling Technique (SMOTE) to generate new records. Merging gene expression datasets with imputation for missing values and integrating cancer type and subtype data could also enhance the model's performance. This approach aims to balance the dataset, enrich the genetic insights, and develop more precise prediction and classification models for EOC recurrence.

Conflicts of Interest

The writers say they have no competing interests.

Author Contributions

The work is done by first author Asha Abraham under the supervision of second author Dr. R. Kayalvizhi and the co-supervision of Dr. S. K. M. Habeeb.

References

[1] R. Aghayousefi,neh, S.M.H. Khatibi, S. Zununi Vahed, M. Bastami, S. Pirmoradi, and M. Teshnehlab, “A diagnostic miRNA panel to detect recurrence of ovarian cancer through artificial intelligence approaches”, *Journal of Cancer Research and Clinical Oncology*, Vol. 149, No. 1, pp.325-341, 2023.

[2] S. Sambasivan, "Epithelial ovarian cancer", *Cancer Treatment and Research Communications*, Vol. 33, pp. 100629, 2022.

[3] https://main.icmr.nic.in/sites/default/files/guidelines/Ovarian_Cancer.pdf

[4] A.K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data", *Neural Computing and Applications*, Vol. 29, pp. 1545-1554, 2018.

[5] I.V. Rodriguez, T. Ghezelayagh, K.P. Pennington, and B.M. Norquist, "Prevention of Ovarian Cancer: Where are We Now and Where are We Going?", *Current Oncology Reports*, pp. 1-12, 2024.

[6] P.K. Illa, S.T. Kumar, and F.S.A. Hussainy, "Deep Learning Methods for Lung Cancer Nodule Classification: A Survey", *J. Mobile Multimedia*, Vol. 18, No. 2, pp. 421-450.

[7] S. Sindhu, D. Hemavathi, K. Sornalakshmi, G. Sujatha, and S. Srividhya, "A Comprehensive Study on the Application of Machine Learning Algorithms in the Prognosis of Ovarian Cancer", *The Open Biomedical Engineering Journal*, Vol. 17, No. 1, 2023.

[8] R.M. Wadapurkar, A. Sivaram, and R. Vyas, "RNA-Seq analysis of clinical samples from TCGA reveal molecular signatures for ovarian cancer", *Cancer Investigation*, Vol. 41, No. 4, pp. 394-404, 2023.

[9] L. A. Corchete, E. A. Rojas, D.A. López, J.D.L. C. Norma, C. Gutiérrez, and F. J. Burguillo, "Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis", *Scientific Reports*, Vol. 10, No. 1, pp. 19737, 2020.

[10] A. Negi, A. Shukla, A. Jaiswar, J.Shrinet, and R. S. Jasrotia, "Applications and challenges of microarray and RNA-sequencing", *Bioinformatics*, pp. 91-103, 2022.

[11] A. Alluri, J. Juneja, C. Khongsai, A. Mishra, and R. K. Gutti, "Identification of Mirs Regulating Oncogenes and Tumour Suppressor Genes in Aml: A Bioinformatic Approach", *SSRN 4914109*, 2024.

[12] P. R. Bushel, S. S. Ferguson, S. C. Ramaiahgari, R. S. Paules, and S. S. Auerbach, "Comparison of Normalization Methods for Analysis of TempO-Seq Targeted RNA Sequencing Data", *Frontiers in Genetics*, Vol. 11, pp. 943, 2020.

[13] N. Gour and P. Khanna, "Ocular diseases classification using a lightweight CNN and class weight balancing on OCT images", *Multimedia Tools and Applications*, Vol. 81, pp. 41765-41780, 2022.

[14] M. Wen and E. B. Tadmor, "Uncertainty quantification in molecular simulations with dropout neural network potentials", *npj computational materials*, Vol. 6, No. 1, pp. 124, 2020.

[15] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review", *Bioengineering*, Vol. 10, No. 2, pp. 173, 2023.

[16] B.H. Kim, K. Yu, and P.C. Lee. "Cancer classification of single-cell gene expression data by neural network," *Bioinformatics*, Vol. 36,No. 5, pp. 1360-1366, 2020.

[17] Q. Rong, W. Wu, Z. Lu, and S. Liao, "Decision-level fusion classification of ovarian CT benign and malignant tumors based on radiomics and deep learning of dual views", *IEEE Access*, 2024.

[18] P. H. Nagarajan and N. Tajunisha, "Automatic Classification of Ovarian Cancer Types from CT Images Using Deep Semi-Supervised Generative Learning and Convolutional Neural Network", *Revue d'Intelligence Artificielle*, Vol. 35, No. 4, 2021.

[19] Y. Zhao, Z. Pan, S. Namburi, A. Pattison, A. Posner, S. Balachander, A.C. Paisie, et al, "CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence", *EBioMedicine*, Vol. 61, 2020.

[20] A. Laios, A. Katsenou, Y.S.Tan, R. Johnson, M. Otify, A. Kaufmann, et al, "Feature selection is critical for 2-year prognosis in advanced stage high grade serous ovarian cancer by using machine learning", *Cancer Control*, Vol. 28, pp. 10732748211044678, 2021.

[21] C. Yan, K. Li, F. Meng, L. Chen, J. Zhao, Z. Zhang, D. Xu, J. Sun, and M. Zhou, "Integrated immunogenomic analysis of single-cell and bulk tissue transcriptome profiling unravels a macrophage activation paradigm associated with immunologically and clinically distinct behaviors in ovarian cancer", *Journal of Advanced Research*, Vol. 44, pp. 149-160, 2023.

[22] https://portal.gdc.cancer.gov/analysis_page?app=Downloads.

[23] J. Lever, E. Y. Zhao, J. Grewal, M. R. Jones, and S. J. Jones, "CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer", *Nature methods*, Vol. 16, No. 6, pp. 505-507, 2019.