IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 September 2023, accepted 24 October 2023, date of publication 1 November 2023,
date of current version 7 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329139

RESEARCH ARTICLE

A Tabular Variational Auto Encoder-Based Hybrid Model for Imbalanced Data Classification With Feature Selection

ASHA ABRAHAM[1], HABEEB SHAIK MOHIDEEN[2], AND R. KAYALVIZHI[1], (Member, IEEE)
[1]Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India
[2]Department of Genetic Engineering College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India

Corresponding author: R. Kayalvizhi (kayalvir@srmist.edu.in)

ABSTRACT Cancer is the deadliest disease in humankind. Ovarian Cancer (OC) is important among female-specific cancers. Epithelial Ovarian Cancer (EOC) is the most commonly occurring subtype of OC. The disease is identified in later stages due to the unrevealed symptoms in the early stages. Gene Expression experiments and machine learning (ML) methodologies can lead to preventive care of OC. This can be achieved by identifying malignant gene transformations earlier and using precision medicine that aids in fast recovery. The proposed hybrid Tabular Variational Auto Encoder oriented dictionary based Stratified K Fold Cross Validation (TVAE_dict_SKCV) is an effective model to handle the threat. The main objective is to assess the significance of EOC screening variables for categorizing high-risk patients. It initially generated synthetic data using the TVAE model to increase the EOC subtype data size from the Cancer Cell Line Encyclopedia. The synthesized data were balanced utilizing the Synthetic Minority Oversampling Technique. Significant features were selected with the Boruta Feature Selection method. The HYPERPARAMETERS were fine-tuned employing Optuna optimizer and applied enhanced SKCV with Random Forest classifier. The TVAE_dict_SKCV method with Boruta acquired an accuracy of 98.5 % and outperformed the experiment with Lasso Feature Selection and with original data. Shapley Additive explanations summarize the main features which classify. Optuna efficiently reduced the computing time compared to the Grid Search Cross Validation optimizer.

INDEX TERMS Machine learning, ovarian cancer, pickle, Optuna, TVAE, Boruta, Lasso.

I. INTRODUCTION

Ovarian cancer (OC) is the development of cancerous cells in the ovaries of the female reproductive system, mostly not seen in early stages. Fast recovery solutions are required for controlling OC. The cells can penetrate and obliterate healthy biological tissue and reproduce swiftly. OC plays a significant role in female cancers. Despite many symptoms, it is often diagnosed late. Therefore, the death rate is higher among those who get this disease. Concerning a study by the American Cancer Society, 1.28 % of women are affected by ovarian cancer. Among them, 0.9 % of patients die. The total duration of treatment for each patient varies, and it depends on the root cause of the illness. The root cause can be found by tracking the genetic variations. Molecular studies on RNA-seq technology play to aid in detecting the genotype factor behind it [1], [2], [3], [4], [5].

EOCs, germ cell tumors, and stromal cell tumors are the three varieties of ovarian cancer. More than 80% of OCs are epithelial ovarian. In females, High-Grade Serous OC (HGSOC), Low-Grade Serous OC (LGSOC), Endometrioid carcinomas, mucinous carcinomas, and clear cell carcinomas are the subtypes of epithelial ovarian type. HGSOC is the most commonly seen among the subtypes [3]. The major problems in molecular studies are less data size and more features. Due to the minimum number of samples, underfitting may occur, and due to the high dimensionality of the data, the issue of overfitting will also appear. Because the dimensions

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.

are given differently in each data set, there are also issues in combining data sets. Machine and Deep learning grounded scrutiny of genes, their expressions, and disease subtypes can enhance precision medicine-related research. ML and DL algorithms can automatically learn the multi-scale attributes to detect the differentially exhibited genes [6], [7], [8].

The implementation unveils the novel TVAE_dict_SKCV model, a hybrid model combining the Tabular Variational Auto Encoder (TVAE) [9] data synthetic method to increase the number of EOC data along with Synthetic Minority Oversampling Technique (SMOTE) for data balancing [10], Boruta Feature Selection (FS) [11], Optuna optimizer [12] for fine-tuning hyperparameters, and dictionary-based SKCV [13] with Random Forest (RF) classifier in classifying the EOC subtypes and feature selection as an early identification mechanism for OC. The selected model was saved with the pickle tool [5], preserving the parameters. The main features that are responsible for classification were recognized with the Shapley Additive explanations (SHAP) [14].

A. MOTIVATION

The limited sample size and the enormous mass of Gene Expression (GE) data features pose significant challenges for molecular cancer research. Along with this, late detection of the cancer disease due to the absence of premature indications exposes the criticality of early malignancy screening. This exhibits how crucial it is to use data synthesizing and feature selection methods to select essential attributes from thousands of features. Considering this, the proposed methodology extracted the differentially expressed (DE) features from the dataset. The extracted features indicate that similar features may be DE in patients.

Regarding the objective, the paper deals with classifying and selecting features of EOC subtypes. The purpose is to discuss the methods involved in developing automatic feature recognition systems, with a view on the underlying concepts, the present literature, and the future perspectives, and having EOC subtypes and features as a potential target. The overview in the article includes the proposed methodology, Experimental analysis, Performance Analysis, Conclusion, and Future enhancement. All the methods used are discussed in the proposed methodology part.

II. LITERATURE SURVEY

Fewer data in molecular datasets is a big challenge in performing research. Data synthesis can be applied as a remedy for increasing sample size. Inan MS et al. [15] created high-quality synthetic tabular data of breast cancers using the conditional generative adversarial network (CTGAN) and Tabular Variational Auto Encoder (TVAE). The classification with TVAE data outperformed CTGAN, achieving 82.83 % accuracy for prognosis and 96.66 % for diagnosis datasets. The presence of unbalanced classes can give biased results. Therefore, balancing the class data is essential. SMOTE is one such technique that can be applied to balance the data among different classes. Ishaq et al. [10] have used SMOTE to handle class unbalancing in identifying survivors of heart failure, achieving an accuracy of 92.6 %. Sorayaie Azar et al. [16] also applied SMOTE to balance OC survival-related classes, fine-tuned data using Grid search cross-validation, classified the categories using an RF classifier, and analyzed significant features using SHAP.

Molecular datasets usually have thousands of features or genes. It is better to choose feature selection algorithms to obtain prominent features. Hwang et al. [17] have applied the Least Absolute Shrinkage and Selection Operator (Lasso) and RF to recognize bone marrow disease from images. The Lasso-RF model acquired a recall percentage of 87.3 and a specificity of 86.2 %, outperforming the Principal Component Analysis Logistic Regression models. Similarly, Casiraghi et al. [11] have used Boruta and RF to predict the variables for coronavirus-infected patients and outperformed other models. Wang and Wang [18] proposed Post-Selection Boosting Random Forest (PBRF) that used Lasso Regression and RF in real-time data analysis. Phung et al. [19] used Boruta for feature selection and RF for classification to identify environmental variables responsible for the deaths of children below five. Htun et al. [20] has done a survey on different Feature selection and feature extraction techniques and explained different types of features considered for the models.

The Optuna optimizer could reduce the time complexity in parameter optimization. Akiba et al. [12] proposed the Optuna Framework, a hyperparameter optimizer. This easily defined lightweight computational method speeds up, parallelizes, and efficiently chooses the best parameters for machine learning algorithms. Using efficient sampling algorithms [TPE, CMA-ES] and pruning algorithms Hyperband] supports the attainment cost and time efficacy. Agrawal [21] created an automated machine learning tool to generate improved ML pipelines on Optuna optimizer. Srinivas and Katarya [22] predicted heart disease by Optuna optimized XGBoost classifier. Hyperparameter Optimization makes the ML and Deep Learning Models more efficient. Shinde et al. [23] utilized a dataset from the UCI ML repository to model various classifiers in investigating liver patients. The proposed system uses Grid search cross-validation (GsCV) for parameter optimization and DT, NB, RF, plus SVM for classifying. RF classifier performed well with accuracy, F1 score, and recall of 72 %, 76.22 %, and 77.37 %.

Class-wise division of data for train-test partitions delivers advanced improvement in performance attainment. This is detailed by El-Gawady et al. [24], whoever preprocessed the GE of Alzheimer's disease (AD) and further split using SKCV and then grouped the classes as AD and normal procuring 97 %, 97 %, 98 %, 98 % for sensitivity, specificity, P, and Acc, approximately. Prusty et al. [13] also explored four usual examination methods by operating SVM, RF, KNN, and XGB classifiers and later applied SKCV. Their findings have shown the RF classifier as a satisfactory replacement, achieving the 95-98 % range for all the tests.

A different methodology is given by Fazelabdolabadi [25] who predicted crude oil price employing a hybrid Bayesian Network. A recent study on OC [26] used the LASSO method to identify differentially expressed prognosis-related genes from genomic data-related computer tomography images. Another work [27] identified biomarkers based on microRNAs by applying the Boruta algorithm and ML algorithms.

The motivation from the literature survey to consider synthetic data generation technique TVAE [15], feature selection techniques [Lasso and Boruta [17], [18]], handling imbalance data by SMOTE [19], fine-tuning the parameters with Optuna [21], classifying the data with SKCV with RF classifier, and projecting the features with SHAP brought out a novel methodology to sub organize EOC data.

III. PROPOSED METHODOLOGY

The main objective of this research paper is to propose a methodology to classify the EOC subtypes and identify the biomarkers responsible for grouping. The outcome of the classification can help in diagnosing disease to check for the differential expression of corresponding biomarkers as an aid for stratified medicine. The proposed methodology architecture and algorithm are shown in Fig. 1, 2.

The proposed methodology contains seven steps: data collection, filtering, generating synthetic data, feature selection, balancing, optimized classification, and a SHAP summary plot. Ovarian genes were filtered from EOC subtype data (HGSOC, LGSOC, endometrioid, clear cell, and mucinous carcinomas) from the Cancer Cell Line Encyclopedia (CCLE) database. The sample size was increased with the help of a TVAE data synthesizer. The synthesized data were balanced by oversampling the data with SMOTE. Once the SMOTE technique is applied, Lasso and Boruta-based FS can be used. The hyperparameters were fine-tuned with an Optuna optimizer for each FS selected feature, and classification was performed in a dict_SKCV RF classifier (Based on split value, assign train-test records for different folds of corresponding split separately). According to train-test records, train the model, perform classification, and store corresponding train-test documents and performance measure values in dictionaries and lists. Finally, significant features of the best model were plotted employing SHAP.

A. DATA SYNTHESIS

Synthetic data is widely used in the financial industry, particularly for risk management, credit risk examination, and fraud detection. Similarly, the TVAE method is applied for synthetic data generation of gene expression data. It is implemented using the Synthetic Data Vault (SDV) package. The model studies the motifs of actual information and gives rise to artificial data. Once data is created, the generated data will be compared with actual samples. If the quality score is above 80 %, it is good. To use Python for this, first build a TVAE model utilizing the tabular package of the SDV tool. A TVAE instance must be used to train the sample data. Later, using that TVAE model, synthetic information can be generated by specifying the total amount of rows required. Additional quality scores can be assessed. In TVAE, the deep generative model, Variational Auto Encoder (VAE), is applied for generating synthetic data from tabular data. In VAE, a regularization term is added over the latent space of the auto-encoder by adding a loss function to avoid overfitting [28]. TVAE is trained using Adam with a learning rate of '1e – 3'. The loss function used in TVAE is Evidence Lower Bound Loss (ELBO). The created artificial data, A(x), can be kept as in Eq. (1).

$$A\left(x\right) = B\left(\text{Decomp}\left(\text{Comp}\left(x\right)\right)\right) \qquad (1)$$

where x represents actual EOC data, B is the TVAE method with x as the entered values and generates A (x). The Comp method, which acts as an Encoder, masters the latent

```
TVAE_dict_SKCV Algorithm:
Input:    Dataset
Output:   EOC subtypes, main features
Process:
Steps:    TVAE_dict_SKCV ()
          {
1           Load EOC GE data
2           Filter out Ovarian genes from Dataset
3           Generate synthetic data using TVAE
4           Apply SMOTE to balance data
5           Perform Boruta FS of balanced data
6.1       Perform train-test split based on split value
6.2       Perform hyperparameter optimization of the
          classifier using Optuna
6.3       Call Stratified k fold (SKCV) with split=3
6.4       Declare lists for storing performance measures
6.5       Declare dictionary for storing train, test sets, and
          trained model based on folds in Split
6.6       For i=0 to split, step by 1, train (0, n, split), test (0, n,
          split)
6.6.1             Assign records for test and train
6.6.2             Store test, train set, and trained model in the
                  dictionary
6.6.3             Perform training based on RF
6.6.4             Predict targets (for training and testing data)
                  based on RF
6.6.5             Calculate accuracy, precision, recall, F1
                  score, and F2 score
6.6.6             Append scores to corresponding lists
6.6.7             Transform train and test target using label
                  encoder method
6.6.8             Calculate MSE for the model
6.6.9             Append scores to corresponding lists
6.7               Repeat Steps 5-11 for different split values
                  of the classifier
7                 Save model with Pickle
          }
```

FIGURE 1. The TVAE_dict_SKCV Algorithm.

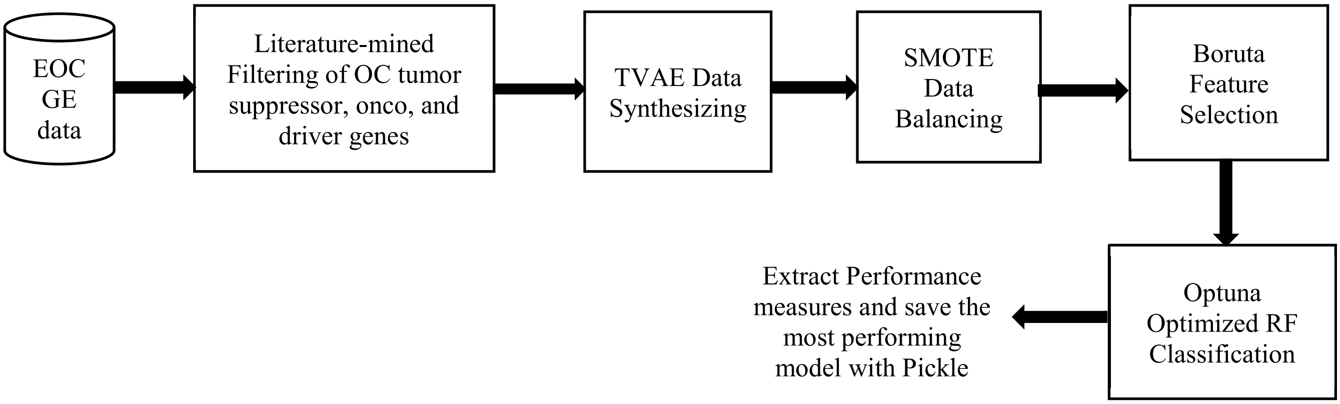
EOC GE data
Literature-mined Filtering of OC tumor suppressor, onco, and driver genes
TVAE Data Synthesizing
SMOTE Data Balancing
Boruta Feature Selection
Optuna Optimized RF Classification
Extract Performance measures and save the most performing model with Pickle

FIGURE 2. **Proposed methodology architecture.**

diffusions of actual data. Later, the Decomp method, the Decoder, generates synthetic data by inspecting the latent diffusions [15].

B. DATA BALANCING

The biases arising from the imbalance between majority and minority classes were avoided by applying the SMOTE technique. The SMOTE method up-samples the minority classes to avoid overfitting. It accomplishes this by generating fresh synthetic examples near other points (of minority class) in the feature space. New neighbors for the minority class were derived based on the Euclidean distance between samples. The sampling rate is proportional to the imbalance in data. New samples' for minority class group R can be generated utilizing Eq. (2).

$$s' = s + \text{rand}\,(0, 1) \times |s - s_I| \tag{2}$$

where s and s_I represent the sample and corresponding neighbors in R [11].

C. FEATURE SELECTION

The number of features can be reduced through FS algorithms. Lasso is on such an algorithm, which is useful when the number of samples is less and the number of elements is high. Lasso reduces the distance between the actual value and predicted value by adjusting the tuning parameter λ so that coefficients of maximum parameters reach near zero. Variables with zero coefficients will be removed from the model. The C_1 regularization term of Lasso, which controls the number of parameters, can be evaluated as in Eq. (3)

$$C_1 = \lambda \times (|a_1| + |a_2| + \ldots |a_d|) \tag{3}$$

wherea_1, a_2, a_d are the coefficients of the parameters [10].

Another similar FS algorithm is the Boruta algorithm. Boruta works with the principle of creating shadow attributes. New attributes can be made by arbitrarily rearranging the existing attributes. Newly generated ones will be merged with currently available ones. The new dataset has to be executed with an RF classifier. Among the old attributes, those with more feature importance than the most important new attribute will be reserved for the ultimate list of attributes.

D. HYPERPARAMETER OPTIMIZATION

The execution of data without parameter optimization may generate poor outcomes. So, it is always good to fine-tune the values of the hyperparameters of the classifier before performing classification. Hyperparameters for the RF classifier were evaluated with the GsCV hyperparameter optimizer. The GsCV creates as many models based on the combination of the number of parameters chosen for the model and the optional values for each parameter. For example, the parameters of RF estimators like n_estimators with weights (10, 20, 30, etc.), max_depth = (5, 10, etc.). Due to the trial with each combination of parameters for each of the algorithm, all possible type of models was trialed. Thus, GsCV takes much computing power and time for execution.

The total number of models, the number of models for the classifier, and the number of fits for a particular test size of the data using GsCV can be evaluated based on the following calculations.

In Eq. (4), 'm' represents each of the parameters for the classifier, and 'n' represents the no of choices for each; the total no of models for the classifier that may be created inside a GsCV can be evaluated as

$$\mathrm{T_{Mc}} = \prod\nolimits_{\mathrm{i=1}}^{\mathrm{N}} m_i n_i \tag{4}$$

If cross-validation, cv= 'p,' with 'p' fits for each of the $\mathrm{T_M}$ models for each test size, the total number of fits $\mathrm{T_F}$ can be evaluated as in Eq. (5)

$$\mathrm{T_F} = pT_{Mc} \tag{5}$$

To save computing power and time of execution of the same work, optimization is trialed with selected hyperparameters using the Optuna hyperparameter optimizer. Optuna optimizer has many sampling and pruning algorithms to remember previously well-performed executions and to stop poor executions early. Optuna's study object is created with direction as maximizer to maximize accuracy, pruner as

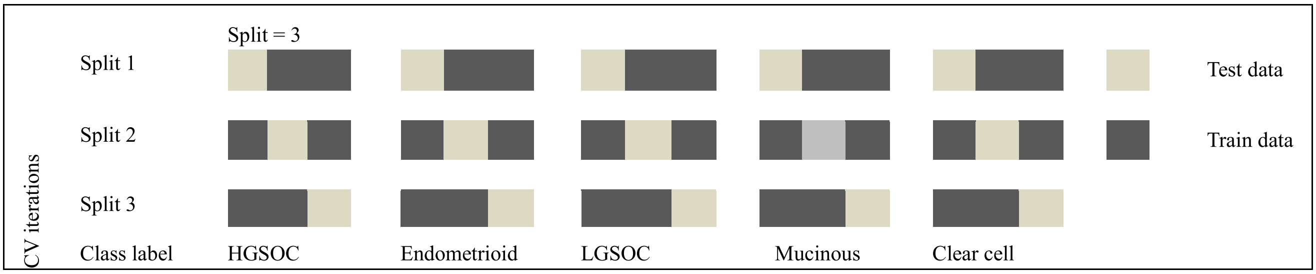

FIGURE 3. The stratified k fold cross validation.

Hyperband algorithm for early stopping, and sampler as TPESampler algorithm. The study object's optimize function was called with objective function and trials as parameters. Tests specify how many times the objective function has to be executed.

The objective function is defined with trial objects to create the model. The trial object suggests parameters specifying whether the parameter's choices are integers, float, categorical, etc. When the algorithm model is defined, the trial-defined variables for parameters have to be assigned to the corresponding algorithm's parameters. The described model has to be trained with the train data from which the score can be evaluated and returned to the study object. The trial suggestions for RF given were int for ['n_estimators,' 'max_depth,'], categorical for ['criterion,' 'bootstrap']. The GsCV's and Optuna's best parameters accomplished the same performance measures, but Optuna efficiently preserved time and computing power [12], [23].

The time complexity of the Optuna optimizer T_O, when the sampler is considered as TPESampler [11], is given in (Eq. (6)), where d stands for search space dimension, and q is the completed trial count.

$$T_O = O(dq(\log q)) \tag{6}$$

E. DICT_SKCV STRATIFIED VALIDATION

Stratified K Fold Cross Validation-based execution of the model can give different combinations of inputs for the train and test set, which can provide variations in performance measures. Consider that the split =3, then one of the three parts of each subtype data will be assigned as test data in the first run. The remaining data will be taken as train data. The next part is given for the test in the second run, the rest as the train. In the third run, the third part will be given for the test section, balance as in Fig. 3. Likewise is the execution when the split takes each assessment 5, 7, and 10, respectively. Final scores will be the mean of both executions plus or minus the standard deviation.

Suppose k acts as the split value {3, 5, 7 or 10}, then the volume of data $\boldsymbol{S_d}$ in similar groups in each fold for a particular split is given in Eq. (7) as

$$\boldsymbol{S_d} = \frac{\mathbf{100}}{\boldsymbol{k}} \tag{7}$$

If $\boldsymbol{C_i}$ can be taken as the count of HGSOC, LGSOC, endometrioid, mucinous, and clear cell individually, then $\boldsymbol{T_t}$, the number of test data for each run (fold or model) in split k, can be derived as in Eq. (8)

$$\boldsymbol{T_t} = \sum\nolimits_{i=1}^{5} \boldsymbol{S_d} \times \boldsymbol{C_i} \tag{8}$$

Similarly, $\boldsymbol{T_r}$ is the number of train data for each run of a particular split can be derived as in Eq. (9)

$$\boldsymbol{T_r} = \boldsymbol{T} - \boldsymbol{T_t} \tag{9}$$

where T is the count of the total data.

In the dict_SKCV model, during each fold execution of a split, the corresponding train and test set can be stored in a dictionary. The performance measures of each fold model can be stored in related lists for later access. With the fine-tuned fold model algorithm, the best-performing model for a split can be evaluated. Based on the fold number, the corresponding stored train, test set, and trained model can be retrieved from the dictionary for the SHAP summary plot of the main contributing features. The fine-tuned FS model algorithm can be applied to choose the best models from all splits depending on the F2 score and mse value once the best fold model from individual split executions is selected. The fine-tuned FS model algorithm can also be applied to choose the best model from both the FS-based executions once the corresponding best model from all splits of individual FS-related executions is completed.

F. FINE-TUNED FOLD MODEL ALGORITHM

To identify the best fold of a specific split of a classifier for input, assume accuracy (Acc) and recall (Rcl) have similar values. If n is the number and split [] the list of splits, P [] the list of precisions, F1 [] the list of F1, R [] the list of Rcl's for all the splits, Ri the Rcl for ith split, then the fold k is chosen for each split based on Fig. 4. For various splits of the classifier, choose model's fold with the leading Rcl value. Compare the absolute difference between the current and prior F values with the current Rcl value when more than one fold element seems to have an identical most considerable Rcl value. Choose the fold number with the most enormous difference. Compare the absolute difference between the current and prior precision (Pr) value with the recent recall (Rcl) value when more than one fold value does have a similar most

```
Fine-tuned fold model Algorithm:
Input  :   All fold for specific split
Output:    The Best fold of a specific split
Process:
           Assign R1 to Rt
           Assign 1 to k
           Repeat n-1 times
               if Rt is less than Ri
                   Assign Ri to Rt
                   Assign i to k
               else if Rt equals Ri
                   if F1k equals F1i
                       then if Pk equals Pi
                               Assign i to k
                       else if absolute (Pk - Ri) is greater
                       than absolute (Pi - Ri)
                               Assign i to k
                   else if absolute (F1k - Ri) greater than
                   absolute (F1i - Ri)
                               Assign i to k
           Select kth fold
```

FIGURE 4. Fine-tuned fold model algorithm.

incredible F value. Pick the fold value well with the most remarkable difference next.

G. FINE-TUNED FS MODEL ALGORITHM

To choose the best model among j no of splits for one particular FS-based classification and then among the best of each FS execution, F2 score (F2) and mse were considered. If M [] is the list of models, F2 [] is the list of F2, mse [] is the list of mse for all the models, F2i is the F2 for the ith model, then the model m is chosen for a particular FS based on Fig. 5. Best models from each FS based classification for the input. Compare the models and, if any, pick the one with the highest F2 value. Choose the model with a lower mse score if many models have the identical most outstanding F2.

```
Fine-tuned FS model Algorithm:
Input:    All best models of each split for specific   FS or
          best models of each FS
Output: Best model among both FS
Process:
        Assign F21 to F2m
            Assign 1 to m
            Repeat j-1 times from i initialized to 2
          If F2m is less than F2i
            Assign F2i to F2m
            Assign i to m
          else if F2m equals F2i
            if msem is greater than msei
                Assign i to m
            Return mth model
```

FIGURE 5. Fine-tuned FS model algorithm.

H. PERFORMANCE MEASURES

The decision about which model to consider can be taken based on performance measures. The key performance indicators used by the work are Acc, Pr, Rcl, $F_{\mathbf{1}}$ score (F**1**), $F_{\mathbf{2}}$, and Mean squared error (mse). They are evaluated depending on the notions of False Negatives (FN), False Positives (FP), True Negatives (TN), and True Positives (TP). The appropriately recognized estimate for each class is the TP. The properly rejected prediction of a group is the TN. FP means wrongly picked out calculations for a group. FN means mistakenly excluded data for a class. The sum of precise forecasts divided by the total quantity of the databank is the Acc (Eq. (10)). In case of imbalanced data with many dissimilarities among different class counts, Acc may mislead. Under such scenarios, low Acc may give a good prediction for all classes compared to high Acc. The F-measure (F) strikes a good compromise between Pr and Rcl. When there is a moderate or substantial disparity between two groups, Pr - Rcl becomes beneficial. i.e., the emphasis must be on accurately classifying the minority class. In medical situations, Rcl is crucial because it should not miss the TP instances while not caring if we trigger a false alert. A measure of a model's performance is its Rcl or sensitivity, defined as the fraction of TP predictions out of all TP predictions (Eq. (11)). With an Rcl of 0.75, the system correctly predicted 75 % of the TPs. Furthermore, Pr or positive predictive value accounts for the correctness of predictions (Eq. 12). F_1 is a harmonic mean of Pr and Rcl (Eq. (13)). F_1 works well for imbalanced data. When a prediction about a class turns out to be TP, it indicates the forecast was accurate. An FP is a prediction of a type that did not occur. Proper identification of ''not-there'' as ''not-there'' is a TN. The actual category was incorrectly labeled as a FN. When discussing cancer prognosis, a ''FP'' refers to a healthy individual incorrectly identified as having cancer, whereas a ''FN'' denotes the opposite.

A variant of the F that emphasizes Rcl is beneficial when dipping FN is of more importance than reducing false positives [13], [24]. F-beta score ($F_{\boldsymbol{\beta}}$) is a generalized form of $F_{\mathbf{1}}(\beta\ =\ 1')$. The $F_{\boldsymbol{\beta}}$ is a simplification of the F in which a coefficient named beta (β) is cast-off to adjust the trade-off between Pr and Rcl in determining the harmonic mean (Eq. (14)). $F_{\boldsymbol{\beta}}$ where beta equals '2.0' gives less weight on Pr, more weight on Rcl [29], [30].

$$\mathrm{Acc} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{TN})} \tag{10}$$

$$\mathrm{Rcl} = \frac{\mathrm{TP}}{(\mathrm{FN}+\mathrm{TP})} \tag{11}$$

$$\mathrm{Pr} = \frac{\mathrm{TP}}{(\mathrm{FP}+\mathrm{TP})} \tag{12}$$

$$\mathrm{F}_1 = \frac{2PR}{(\mathrm{P}+\mathrm{R})} \tag{13}$$

$$F_\beta = \frac{\left(1+\beta^2\right) PR}{\left(\beta^2 P + R\right)} \tag{14}$$

Assuming $\boldsymbol{R_{cj}}$ holds the recall values of HGSOC, LGSOC, endometrioid, mucinous, and clear cell, the average *Rcl* for EOC and n=5, i.e., $\mathrm{avgR}_{\boldsymbol{EOC}}$ can be calculated as in Eq. (15)

$$\mathrm{avgR_{EOC}} = \frac{\sum_{j=1}^{n} \boldsymbol{R_{cj}}}{\boldsymbol{n}} \tag{15}$$

Assuming $\boldsymbol{P_{cj}}$ holds the precision values of HGSOC, LGSOC, endometrioid, mucinous, and clear cell, the average *Pr* for EOC, i.e., $avgP_{\boldsymbol{EOC}}$,can be calculated as follows in Eq. (16).

$$\mathrm{avgP_{EOC}} = \frac{\sum_{j=1}^{n} \boldsymbol{P_{cj}}}{\boldsymbol{n}} \tag{16}$$

More priority for Rcl than Pr is given by substituting β as 2 in Eq. (14). Hence, Eq. (14) is replaced with the average values of recall in Eq. (15) and precision in Eq. (16) and $\mathrm{F_2}$ for EOC. i.e., $\mathrm{F_{2EOC}}$ is given in Eq. (17).

$$\mathrm{F_{2EOC}} = \frac{\boldsymbol{n} \times \mathbf{avgP_{EOC}} \times \mathbf{avgR_{EOC}}}{(\boldsymbol{n}-\mathbf{1}) \times \mathbf{avgP_{EOC}} + \mathbf{avgR_{EOC}}} \tag{17}$$

A high deviation of the predicted target value from the actual one or poor performance can cause high bias in the model. The model's variance is increased when there is too much variation between train and test accuracies. The appropriate bias and variance scales control the overfitting and underfitting issues. A model is described as an overfit, underfit, and balanced fit based on a bias-variance trade-off. Class balancing, fine-tuned hyperparameters, and relevant features are required to keep the model proportional to fit adequate samples. Including the mentioned characteristics in the model will improve its performance and reduce error. The model that makes fewer errors generates more accurate predictions [31].

IV. EXPERIMENTAL ANALYSIS

The experiment was initially done by classifying with actual data, including all the available features. Initial performance was improved with grid search parameter optimization, followed by Optuna optimization. For further enhancement, disease-specific markers or genes were selected. Based on the chosen features, classification with the FS methods Boruta and Lasso was conducted to avoid overfitting. Synthetic data were generated with TVAE data synthesizer to prevent underfitting issues and boost achievements. The synthesized data classes were balanced by applying the SMOTE technique. A dictionary and list-oriented stratified k-fold cross-validation of the balanced data were conducted using both FS methods after finding the fine-tuned values for the hyperparameters utilizing Optuna optimization. Optimization was performed after splitting the data into test and train based on t-he split value for SKCV. Ensuing SKCV, based on the performance measures and error value, the best model for each split, the best model for each FS method, and the best FS method are also chosen.

A. DATASET DESCRIPTION

Epithelial OC-based gene expression data is retrieved from CCLE [32], [33] and is sorted to isolate information specific to EOC with its subtypes HGSOC, LGSOC, endometrioid, clear cell, and mucinous carcinomas. Sixteen thousand three hundred eighty-three gene columns with forty-four sample sizes were there before preprocessing. The data gets normalized by removing the gene rows with row sums of zero, less than ten, NAN, and then identical genome fields. The feature size became 13399 after preprocessing. Gene names of the repository act as attributes or table headers.

B. OPTUNA-BASED ML IMPLEMENTATION

Initially, a simple data classification with 13399 features was done with a Random Forest algorithm with an accuracy of 60%. Later, the parameters of RF, especially (n_estimators and max_depth), were fine-tuned initially with Grid search and later with Optuna optimization methods. With both optimization methods, classification performance improved to an accuracy of 86%. However, the execution time of Optuna was much less than the Grid search, as mentioned in Table. 1. So, for the remaining work done with synthesized data and with the data balancing technique, Optuna optimization was chosen for fine-tuning the parameter values of the classifier.

TABLE 1. Time comparisons for parameter optimizations.

	Time in seconds	
Classifier	Grid search	Optuna
RF	331.87	37.81

C. FEATURE SELECTION ON ACTUAL DATA

There are disease-specific pathways for each type of disease. Selecting features based on pathways can lead to correctly classifying and identifying disease-causing features. To facilitate this, 1123 OC-related genes were taken from Cancermine (an article-extracted dataset of cancer drivers, oncogenes, and tumor suppressors, which contains genes associated with malignancy from various OC oncogenic signaling pathways allied articles as well) [34], Cancer Gene Census [35] and National Comprehensive Cancer Network(NCCN) guidelines(24 genes) [36] depending on EOC subtypes [High-Grade Serous OC(HGSOC), endometrioid, Low-Grade Serous OC(LGSOC), clear cell, mucinous, and control data]. Among 1123 OC genes, 941 were present in the EOC data from CCLE.

TABLE 2. FS on actual data.

	Acc	Pr	Rcl	F_1	F_2	mse
Boruta-RF	87.5	78.1	87.5	82.1	85.1	25.6
Lasso-RF	88.8	81.4	88.8	84.4	86.9	42.6

Lasso and Boruta FS were done on the 941 genes of the actual data. RF classification was performed based on

76 features selected by Lasso FS and 200 by the Boruta algorithm. Based on FS, accuracy and recall have improved for both algorithms {87.5 %, 88.8 % for Boruta-RF, Lasso-RF}, but have high error values {25.26 %, 42.6 % for Boruta-RF, Lasso-RF} as given in Table 2, [11], [17]. This is an indication of more sample requirements.

D. TVAE DATA SYNTHESIS, SMOTE DATA BALANCING

Considering the reduction of error and improving performance as the aim, with the aid of the TVAE method from Python's 'sdv.tabular' package generated 200 synthetic data consuming the 44 actual samples. The evaluation quality score of synthetic data with actual data was 99 %. The unbalanced synthetic data were balanced with the SMOTE method, improving to a count of 455 samples.

E. FS AND DICT_SKCV ON SYNTHESIZED DATA

FS applied for the 455 samples originating 220 Lasso and 452 Boruta features. The samples with the Lasso and Boruta selected features underwent dict_SKCV with RF as the classifier. The result of the execution with different split values {3, 5, 7, and 10} is given in Table 3. For both FS-oriented classifications, split 10 has high-performance values. Boruta-based category outperformed Lasso with {Acc, Pr, Rcl, F1, F2, and mse} as {98.5, 98.6, 98.5, 98.5, 98.5, and 06}.

TABLE 3. Performance measures from synthesized data.

Model	Split	Acc	Pr	Rcl	F_1	F_2	mse
TVAESBRF	3	93.4	93.8	93.4	93	93.1	14
TVAESBRF	5	95.6	95.8	95.6	95.5	95.5	12.4
TVAESBRF	7	96.9	97.1	96.9	96.9	96.9	7.9
TVAESBRF	10	98.5	98.6	98.5	98.5	98.5	06
TVAESLRF	3	92	92.5	92	91.5	91.6	16
TVAESLRF	5	95.6	96.1	95.6	95.4	95.5	12.6
TVAESLRF	7	95.4	95.4	95.4	95.3	95.3	10.3
TVAESLRF	10	97.8	98	97.8	97.8	97.7	7.6

* TVAESBRF: = TVAE + SMOTE + Boruta + SKCV_RF
* TVAESLRF: = TVAE + SMOTE + Lasso + SKCV_RF

TABLE 4. Execution time of synthesized data.

Model	Split	Time in minutes
TVAESBRF	3	10
TVAESBRF	5	13
TVAESBRF	7	11
TVAESBRF	10	20
TVAESLRF	3	11
TVAESLRF	5	9
TVAESLRF	7	5
TVAESLRF	10	13

Regarding the execution time, it can be noticed from Table 4 that, generally, computational complexity is a bit complex for Boruta FS classification in comparison to Lasso. As the number of trees increases (n_estimator) or the depth of the tree, a hike in computational time is observed. If either n_estimator or max_depth is a low value or both have some medium value, the execution time will be less. A slight dependency on Google Colab's internet speed was also observed.

The SHAP summary plot, which shows the main features of each class and the class-vise importance level of the features in classification, is in Fig. 6. The top 10 features from the SHAP summary plot of each category are provided in Table 5.

Some article-related information about the genes was noticed in the molecular biology techniques-based articles

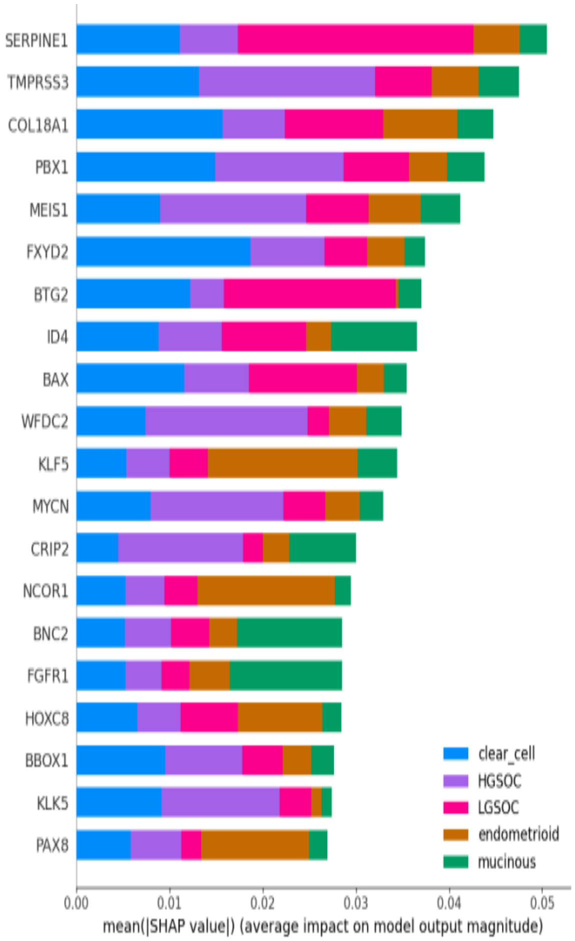

FIGURE 6. Fine-tuned FS model algorithm.

TABLE 5. Top 5 SHAP features of individual subtypes.

Subtype	Features
HGSOC	{EPCAM, ACLY, SFRP1, STC1, MXRA5, PTPRB, ROS1, S100A14, CCL2, TK1}
LGSOC	{ROS1, FSCN1, ZEB2, PADI2, MIEF2, STC1, GADD45B, DGCR8, MUTYH, MDM2}
Clear cell	{NDN, TCF4, NFATC2, MDM2, ZEB2, KDM1A, ID4, ROS1, ZEB1, STC1}
Endometrioid	{ROS1, PTPRB, ZEB2, EPCAM, NDN, ZEB1, MUTYH, TCF4, GPER1, FSCN1}
Mucinous	{ID4, CCL2, ZEB1, SLC1A5, ZEB2, CERS6, NDN, NOTCH2, NFATC2, STC1}

published afterward. In their OC pathway-related studies, Agustriawan et al. [37] stated that the part of NDN in OC poor survival. Tayama et al. [38] verified that EpCAM-positive cells resist cisplatin-related apoptosis and may lead to recurrent disease, though EpCAM-negative cells get eliminated after chemotherapy in EOC cases. Meel et al. [39] wrote about the role of differentially expressed genes ZEB1, ZEB2, and NOTCH in different cancers like OC, breast cancer, etc., in the epithelial to mesenchymal transition (EMT). This process leads to cancer progression as metastasis and chemoresistance.

Dong et al. [40] reported the detection of ROS1 mutation for HGSOC patients, followed by improved recovery after administering Crizotinib treatment. Li et al. [41] conveyed that up-regulation of ZEB2 in HGSOC patients has seen results in the metastasis state of the disease. Wang et al. [42], through experiments with western blotting and RT-qPCR, found the overexpression of FSCN1 and CRNDE in OC cells. Bajwa et al. [43], with their investigations, explained the role of STC1 and ANGPTL4 in initiating OC metastasis.

F. PERFORMANCE ANALYSIS

Performance comparison of baseline and proposed models are given in Table 6. Good performance measure of {98.5, 97.8} percent accuracy, recall achieved for synthetic TVAESBRF, TVAESLRF proposed models (Fig. 7,9). Low mse values of {06, 7.6} were obtained in the case of proposed models (Fig. 11). The result shows that the TVAE_dict_SKCV model performs well and can be used for synthetic data generation of similar small datasets since it generated data with good quality score, less error, high accuracy and having contributing features from molecular biology related articles.

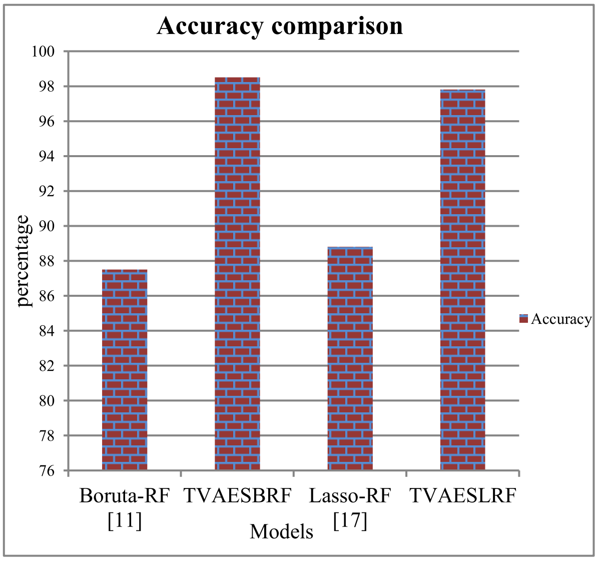

FIGURE 7. Accuracy comparison of actual, synthesized data.

TABLE 6. Performance measures of actual, synthesized data.

	Acc	Pr	Rcl	F_1	F_2	mse
Boruta-RF[11]	87.5	78.1	87.5	82.1	85.1	25.6
TVAESBRF	98.5	98.6	98.5	98.5	98.5	06
Lasso-RF[17]	88.8	81.4	88.8	84.4	86.9	42.6
TVAESLRF	97.8	98	97.8	97.8	97.7	7.6

* TVAESBRF: = TVAE + SMOTE + Boruta +SKCV_RF (Synthetic)
* TVAESLRF: = TVAE + SMOTE + Lasso + SKCV_RF (synthetic)
* Boruta-RF: = (actual data)
* Lasso-RF: = (actual data)

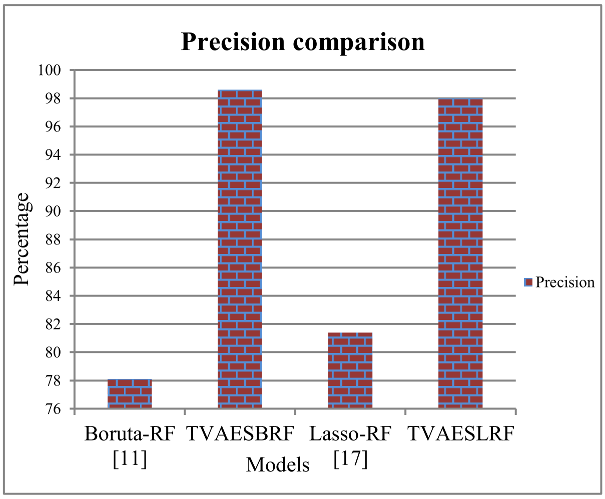

FIGURE 8. Precision comparison of actual, synthesized data.

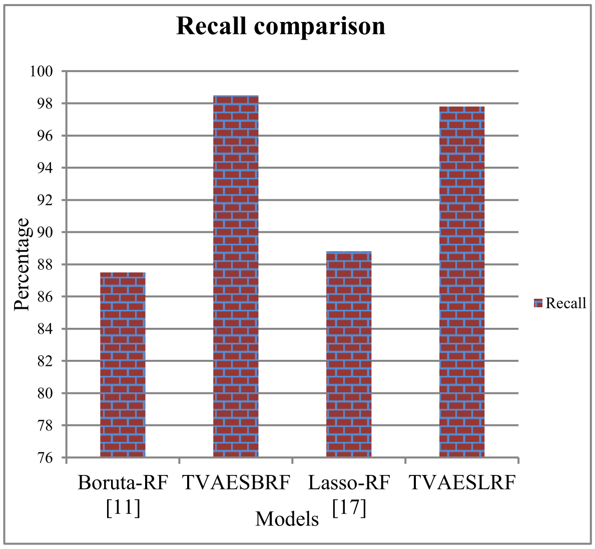

FIGURE 9. Recall the comparison of actual, synthesized data.

Thus, the proposed system achieved improved accuracy, precision, recall, F1 score, F2 score, and MSE of {98.5, 98.6,

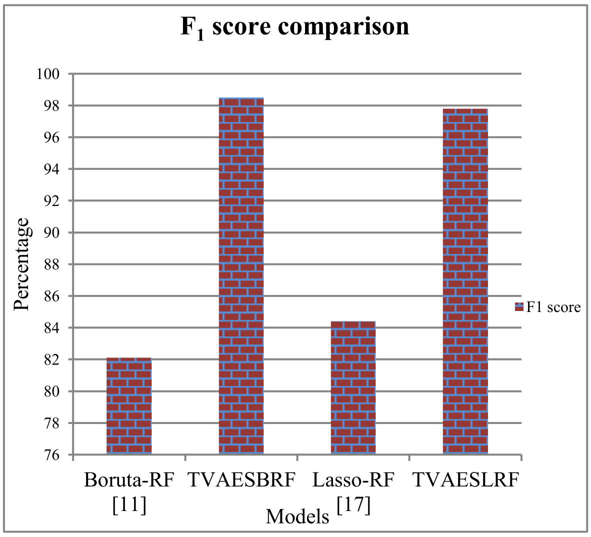

FIGURE 10. F1 score comparison of actual, synthesized data.

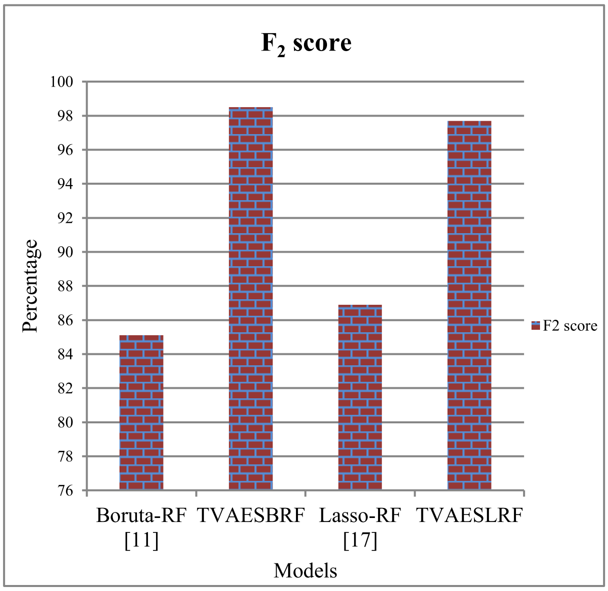

FIGURE 11. F2 score comparison of actual, synthesized data.

98.5, 98.5, 98.5, and 06} percentages with Boruta Feature Selection.

The paper contributes a method to generate more related patterns with good quality scores in case of a shortage of research data. Further, the framework provides a mechanism to avoid bias in identification due to irregularities. The usage of the optimizer drastically improved performance. The FS technique removes insignificant features, thereby narrowing down to the proper classification and interpretations of differentially expressed characteristics.

The presented work can be improved by incorporating some other type of data. The disease's stages, recurrence, and metastasis details can be combined with this work. Other types of cancer information can also be incorporated to find commonly occurring gene transformations and their relationships.

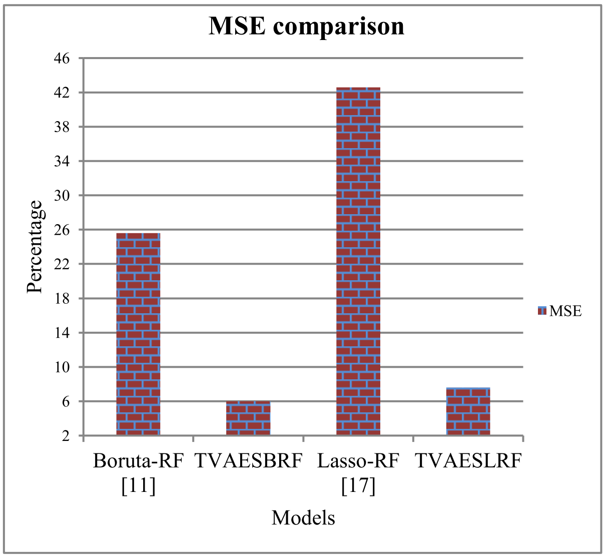

FIGURE 12. MSE comparison of actual, synthesized data.

V. CONCLUSION AND FUTURE ENHANCEMENT

Applying RNA-seq and ML techniques together may pave the road to targeted therapy, which speeds up the healing process by pinpointing the mutated genes that cause cancer. The research is carried out by classifying mRNA-based EOC data from the CCLE into subtypes using the proposed TVAE_dict_SKCV. The usage of the Optuna optimizer improved the performance. Application of TVAE to the actual data produced synthetic data with good quality scores. The new data aided in accomplishing reasonable performance measures of {98.5, 98.6, 98.5, 98.5, 98.5, and 6} % accuracy, precision, recall, F1 score, F2 score, and mse values. Compared to the accuracy of Boruta and Lasso models with actual data {87.5, 88.8} %, accuracy has been improved to {98.5, 97.8} % with TVAE-based synthesized data. The novel method paves the way for increasing data size and finding possible biomarkers for similar data. As the interpretable features from the SHAP summary plot are found to be mentioned in the molecular biology experiments-based articles as biomarkers for OC, TVAE_dict_SKCV can be used for similar other small datasets. In the future, the work can be extended with more data from similar and different disease types. Also, adding standard data to the subtypes can clarify the classification and identification of biomarkers.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare that are relevant to this article's content.

AVAILABILITY OF DATA AND MATERIALS

The data is available from the Cancer Cell Line Encyclopedia website.

FUNDING STATEMENT

The author(s) received no specific funding for the work.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA, A Cancer J. Clinicians*, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/CAAC.21708.

[2] M. R. Radu, A. Prădatu, F. Duică, R. Micu, S. M. Creţoiu, N. Suciu, D. Creţoiu, V. N. Varlas, and V. E. Rădoi, "Ovarian cancer: Biomarkers and targeted therapy," *Biomedicines*, vol. 9, no. 6, p. 693, Jun. 2021.

[3] E. Hulstaert, A. Morlion, K. Levanon, J. Vandesompele, and P. Mestdagh, "Candidate RNA biomarkers in biofluids for early diagnosis of ovarian cancer: A systematic review," *Gynecologic Oncol.*, vol. 160, no. 2, pp. 633–642, 2021.

[4] A. Abraham, R. Kayalvizhi, H. S. Mohideen, and A. Thiyagaraj, "Malignancy transcriptome analysis, tools and deep learning methodologies for prediction of diseases," in *Proc. Int. Conf. Adv. Comput., Commun. Appl. Informat. (ACCAI)*, Jan. 2022, pp. 1–11.

[5] A. Abraham, R. Kayalvizhi, H. S. Mohideen, and A. Abraham, "CWAOMT: Class weight balanced artificial neural network model for the classification of ovarian malignancy from transcriptomic profiles," in *Proc. Int. Conf. Netw. Commun. (ICNWC)*, Apr. 2023, pp. 1–6.

[6] J. Vamathevan, D. Clark, and P. Czodrowski, "Applications of machine learning in drug discovery and development," *Nature Rev. Drug Discovery*, vol. 18, no. 6, pp. 463–477, Apr. 2019.

[7] C. Xu and S. A. Jackson, "Machine learning and complex biological data," *Genome Biol.*, vol. 20, no. 1, pp. 1–4, Dec. 2019.

[8] S. K. Prabhakar and S.-W. Lee, "An integrated approach for ovarian cancer classification with the application of stochastic optimization," *IEEE Access*, vol. 8, pp. 127866–127882, 2020.

[9] J. Rivers, A. Nelson, and L. Williams. *Synthetic Data Generation With SDV*. Accessed: May 2022. [Online]. Available: https://researchgate.net

[10] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021.

[11] E. Casiraghi, D. Malchiodi, G. Trucco, M. Frasca, L. Cappelletti, T. Fontana, A. A. Esposito, E. Avola, A. Jachetti, J. Reese, A. Rizzi, P. N. Robinson, and G. Valentini, "Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments," *IEEE Access*, vol. 8, pp. 196299–196325, 2020.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

[13] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers Nanotechnol.*, vol. 4, Aug. 2022, Art. no. 972421.

[14] Y. Wang and J. Lang, "The radiomic-clinical model using the SHAP method for assessing the treatment response of whole-brain radiotherapy: A multicentric study," *Eur. Radiol.*, vol. 32, no. 12, pp. 8737–87477, 2022.

[15] M. S. K. Inan, S. Hossain, and M. N. Uddin, "Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor's morphological information," *Informat. Med. Unlocked*, vol. 37, 2023, Art. no. 101171.

[16] A. S. Azar, S. B. Rikan, A. Naemi, J. B. Mohasefi, H. Pirnejad, M. B. Mohasefi, and U. K. Wiil, "Application of machine learning techniques for predicting survival in ovarian cancer," *BMC Med. Informat. Decis. Making*, vol. 22, no. 1, p. 345, Dec. 2022.

[17] E.-J. Hwang, S. Kim, and J.-Y. Jung, "Bone marrow radiomics of T1-weighted lumber spinal MRI to identify diffuse hematologic marrow diseases: Comparison with human readings," *IEEE Access*, vol. 8, pp. 133321–133329, 2020.

[18] H. Wang and G. Wang, "Improving random forest algorithm by Lasso method," *J. Stat. Comput. Simul.*, vol. 91, no. 2, pp. 353–367, 2021.

[19] V. L. H. Phung, K. Oka, Y. Hijioka, K. Ueda, M. Sahani, and W. R. Wan Mahiyuddin, "Environmental variable importance for under-five mortality in malaysia: A random forest approach," *Sci. Total Environ.*, vol. 845, Nov. 2022, Art. no. 157312.

[20] H. H. Htun, M. Biehl, and N. Petkov, "Survey of feature selection and extraction techniques for stock market prediction," *Financial Innov.*, vol. 9, no. 1, p. 26, 2023.

[21] T. Agrawal, *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. New York, NY, USA: Apress, 2021.

[22] P. Srinivas and R. Katarya, "HyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103456.

[23] S. Shinde, "Liver disease prediction based on grid search and random forest classification," *Int. J. Eng. Appl. Sci. Technol.*, vol. 7, no. 1, pp. 136–140, May 2022.

[24] A. El-Gawady, M. A. Makhlouf, B. S. Tawfik, and H. Nassar, "Machine learning framework for the prediction of Alzheimer's disease using gene expression data based on efficient gene selection," *Symmetry*, vol. 14, no. 3, p. 491, Feb. 2022.

[25] B. Fazelabdolabadi, "A hybrid Bayesian-network proposition for forecasting the crude oil price," *Financial Innov.*, vol. 5, no. 1, pp. 1–21, 2019.

[26] S. Wan, T. Zhou, R. Che, Y. Li, J. Peng, Y. Wu, S. Gu, J. Cheng, and X. Hua, "CT-based machine learning radiomics predicts CCR5 expression level and survival in ovarian cancer," *J. Ovarian Res.*, vol. 16, no. 1, pp. 1–15, Jan. 2023.

[27] F. Hamidi, N. Gilani, R. Arabi Belaghi, H. Yaghoobi, E. Babaei, P. Sarbakhsh, and J. Malakouti, "Identifying potential circulating miRNA biomarkers for the diagnosis and prediction of ovarian cancer using machine-learning approach: Application of Boruta," *Frontiers Digit. Health*, vol. 5, pp. 1–13, Aug. 2023.

[28] J. Rocca. (Sep. 2019). *Understanding Variational Autoencoders (VAEs) Towards Data Science*. [Online]. Available: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

[29] B. Norgeot, K. Muenzen, T. A. Peterson, X. Fan, B. S. Glicksberg, G. Schenk, E. Rutenberg, B. Oskotsky, M. Sirota, J. Yazdany, G. Schmajuk, D. Ludwig, T. Goldstein, and A. J. Butte, "Protected health information filter (Philter): Accurately and securely de-identifying free-text clinical notes," *NPJ Digit. Med.*, vol. 3, no. 1, p. 57, Apr. 2020.

[30] C. Morais, K. L. Yung, K. Johnson, R. Moura, M. Beer, and E. Patelli, "Identification of human errors and influencing factors: A machine learning approach," *Saf. Sci.*, vol. 146, Feb. 2022, Art. no. 105528.

[31] P. Chakraborty, S. S. Rafiammal, C. Tharini, and D. N. Jamal, "Influence of bias and variance in selection of machine learning classifiers for biomedical applications," in *Smart Data Intelligence*. Singapore: Springer Nature, 2022, pp. 459–472.

[32] M. Ghandi, "Next-generation characterization of the cancer cell line encyclopedia," *Nature*, vol. 569, no. 7757, pp. 503–508, 2019.

[33] B. M. Barnes, L. Nelson, A. Tighe, G. J. Burghel, I.-H. Lin, S. Desai, J. C. McGrail, R. D. Morgan, and S. S. Taylor, "Distinct transcriptional programs stratify ovarian cancer cell lines into the five major histological subtypes," *Genome Med.*, vol. 13, no. 1, pp. 1–9, Dec. 2021.

[34] J. Lever and E. Y. Zhao, "CancerMine: A literature-mined resource for drivers, oncogenes and tumor suppressors in cancer," *Nature Methods*, vol. 16, no. 6, pp. 505–507, 2019.

[35] K. Salokas, R. G. Weldatsadik, and M. Varjosalo, "Human transcription factor and protein kinase gene fusions in human cancer," *Sci. Rep.*, vol. 10, no. 1, p. 14169, Aug. 2020.

[36] D. K. Armstrong and R. D. Alvarez, "NCCN guidelines® insights: Ovarian cancer, version 3.2022: Featured updates to the NCCN guidelines," *J. Nat. Comprehensive Cancer Netw.*, vol. 20, no. 9, pp. 972–980, 2022.

[37] D. Agustriawan and C. H. Huang, "DNA methylation-regulated microRNA pathways in ovarian serous cystadenocarcinoma: A meta-analysis," *Comput. Biol. Chem.*, vol. 65, pp. 154–164, Dec. 2016.

[38] S. Tayama, T. Motohara, D. Narantuya, C. Li, K. Fujimoto, I. Sakaguchi, H. Tashiro, H. Saya, O. Nagano, and H. Katabuchi, "The impact of EpCAM expression on response to chemotherapy and clinical outcomes in patients with epithelial ovarian cancer," *Oncotarget*, vol. 8, no. 27, pp. 44312–44325, Jul. 2017.

[39] M. H. Meel, S. A. Schaper, G. J. Kaspers, and E. Hulleman, "Signaling pathways and mesenchymal transition in pediatric high-grade glioma," *Cellular Mol. Life Sci.*, vol. 75, no. 5, pp. 871–887, 2018.

[40] D. Dong, G. Shen, Y. Da, M. Zhou, G. Yang, M. Yuan, and R. Chen, "Successful treatment of patients with refractory high-grade serous ovarian cancer withGOPC-ROS1 fusion using crizotinib: A case report," *Oncologist*, vol. 25, no. 11, pp. e1720–e1724, Nov. 2020.

[41] Y. Li and H. Fei, "ZEB2 facilitates peritoneal metastasis by regulating the invasiveness and tumorigenesis of cancer stem-like cells in high-grade serous ovarian cancers," *Oncogene*, vol. 40, no. 32, pp. 5131–5141, 2021.

[42] Q. Wang and L. X. Wang, ''LncRNA CRNDE promotes cell proliferation, migration and invasion of ovarian cancer via miR-423–5p/FSCN1 axis,'' *Mol. Cellular Biochemistry*, vol. 477, no. 5, pp. 1477–1488, 2022.

[43] P. Bajwa, K. Kordylewicz, A. Bilecz, R. R. Lastra, K. Wroblewski, Y. Rinkevich, E. Lengyel, and H. A. Kenny, ''Cancer-associated mesothelial cell–derived ANGPTL4 and STC1 promote the early steps of ovarian cancer metastasis,'' *JCI Insight*, vol. 8, no. 6, pp. 1–16, Mar. 2023.

ASHA ABRAHAM received the B.E. degree in computer science and engineering and the M.Tech. degree in information technology. She is currently pursuing the Ph.D. degree in artificial intelligence related to health care data with the Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai. She has experience in the field of teaching. Her research interests include artificial intelligence, bioinformatics, big data, healthcare, the Internet of Things, and computer vision.

HABEEB SHAIK MOHIDEEN received the Ph.D. degree in bioinformatics from the University of Madras, in 2012. He is currently an Associate Professor and a Principal Investigator with the Bioinformatics and Entomoinformatics Laboratory, Department of Genetic Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai. He has an experience of about 14 years in teaching and research. He is a Bioinformatician with expertise in developing and applying applications to solve biological problems. A bioinformatics portal, SRMBIT, developed by him, hosts several bioinformatics tools, primarily a neural network-based secondary structure prediction tool, validating the prediction with the experimentally determined structures. He is working in genomics (NGS) and bioinformatics and has executed government-funded projects on genomics and bioinformatics and is also executing an international collaborator grant. His expertise in RNAseq, small RNAseq, metagenomics, mitochondrial genomics, and exome sequencing and its data analysis has contributed to more than 17 NGS entries into the NCBI SRA database. His work on whole mitochondrial genome sequencing on dusky cotton bugs has been well received and is among the first on this specimen. He has guided two and is also guiding six Ph.D. students. He has several publications in peer-reviewed journals and holds international collaborations to work in genomics and bioinformatics.

R. KAYALVIZHI (Member, IEEE) received the B.E. degree in computer science and engineering from Madras University, in 2000, the M.E. degree in embedded system technologies from the College of Engineering, Anna University, Chennai, in 2007, and the Ph.D. degree in wireless sensor network security from MIT Campus, Anna University, in 2016. She is currently an Assistant Professor with the Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai. Her research interests include cryptography, network security, wireless sensor networks, healthcare, and real-time systems.