*mathematics*

MDPI

*Article*

# An Integrated Algorithm with Feature Selection, Data Augmentation, and XGBoost for Ovarian Cancer

**Jingxun Cai [1], Zne-Jung Lee [2,*], Zhihxian Lin [2,*], Chih-Hung Hsu [3,*] and Yun Lin [4]**

[1] Graduate School of New Generation Electronic Information Engineer, School of Advanced Manufacturing, Fuzhou University, Quanzhou 362200, China; lemoonwwc@gmail.com

[2] Department of Electronic and Information Engineering, School of Advanced Manufacturing, Fuzhou University, Quanzhou 362200, China

[3] Institute of Logistics Engineering and Management, College of Transportation, Fujian University of Technology, Fuzhou 350118, China

[4] School of Intelligent Construction, Fuzhou University of International Studies and Trade, Fuzhou 350200, China; lyun@fzfu.edu.cn

* Correspondence: johnlee@fzu.edu.cn (Z.-J.L.); lzx2005000@163.com or t98068@fzu.edu.cn (Z.L.); chhsu@fjut.edu.cn (C.-H.H.)

**Abstract:** Ovarian cancer is one of the most aggressive gynecological cancers due to its high invasion and chemoresistance. It not only has a high incidence rate but also tops the list of mortality rates. Its subtle early symptoms make subsequent diagnosis difficult, significantly delaying timely treatment for patients. Once ovarian cancer reaches an advanced stage, the complexity and difficulty of treatment increase substantially, affecting patient survival rates. Therefore, it is crucial for both medical professionals and patients to remain highly vigilant about the early signs of ovarian cancer to ensure timely intervention. In recent years, ovarian cancer prediction research has advanced, allowing for the analysis of the likelihood and type of cancer based on patients' genetic data. With the rapid development of machine learning, numerous efficient classification prediction models have emerged. These new technologies offer significant opportunities and potential for developing ovarian cancer diagnostic prediction methods. However, traditional approaches often struggle to achieve satisfactory classification accuracy in high-dimensional genetic datasets with small sample sizes. This research offers a prediction model utilizing genomic data to enhance the early diagnosis rate of ovarian cancer, incorporating feature selection, data augmentation through adversarial conditional generative adversarial networks (AC-GAN), and an extreme gradient boosting (XGBoost) classifier. First, we can simplify the original genetic dataset through feature selection methods, removing irrelevant variables and noise, thereby improving the model's predictive accuracy. Following dimensionality reduction, AC-GAN enriches the data, producing more realistic genetic samples to enhance the model's generalization capacity. Finally, the XGBoost classifier is applied to classify the augmented data, achieving efficient predictions for ovarian cancer. These research findings strongly demonstrate that the diagnostic method proposed in this paper has a significant advantage in the predictive diagnosis of ovarian cancer, with an accuracy of 99.01% that surpasses the current technologies in use. Additionally, the algorithm identifies twelve genes highly relevant to ovarian cancer, providing valuable insights for physicians during diagnosis.

**Keywords:** integrated algorithm; ovarian cancer; feature selection; data augmentation; extreme gradient boosting; generative adversarial networks

**MSC:** 92-08

check for updates

**Citation:** Cai, J.; Lee, Z.-J.; Lin, Z.; Hsu, C.-H.; Lin, Y. An Integrated Algorithm with Feature Selection, Data Augmentation, and XGBoost for Ovarian Cancer. *Mathematics* **2024**, *12*, 4041. https://doi.org/10.3390/math12244041

Academic Editor: Konstantin Kozlov

Received: 11 November 2024
Revised: 19 December 2024
Accepted: 22 December 2024
Published: 23 December 2024

## 1. Introduction

Ovarian cancer is a malignant tumor in women's ovaries, ranking fifth in cancer-related mortality among women. According to GLOBOCAN 2022, the global, age-adjusted incidence rate is seven cases per 100,000 women, with a similar fatality rate of six deaths per 100,000 women worldwide [1]. Although ovarian cancer can affect women of any age, most diagnoses occur after age 55 [2]. Unfortunately, only a quarter of patients are diagnosed at an early stage, as symptoms are often subtle and may be misinterpreted as less serious conditions, like indigestion or irregular menstrual cycles. Early diagnosis is critical when the tumor is still confined to the ovaries, as it improves prognosis and increases cure rates, while also reducing the complexity and suffering of treatment [3].

Early diagnosis of ovarian cancer has always been a medical challenge, mainly because we currently lack screening methods that are sensitive and specific enough to effectively prevent and detect this disease at an early stage. It is clear that existing screening methods, such as CA-125 blood tests and transvaginal ultrasound examinations, help to identify ovarian cancer to some extent. However, their accuracy is not high, and they are prone to false positives and false negatives [4]. It is, therefore, clear that current medical guidelines do not recommend routine ovarian cancer screening for asymptomatic women. The scientific evidence is currently insufficient to prove that these screening methods can significantly reduce the mortality risk of ovarian cancer in the general population [5]. Women with a family history of ovarian or other cancers, who may have inherited susceptibility genes, can be regularly monitored through specific biomarkers in blood tests or various imaging examinations to detect early signs of cancer. Some positive results have been observed [6]. In the 1980s, Bast and his colleagues discovered that the carbohydrate antigen CA125 was a crucial diagnostic for ovarian cancer detection [7]. However, conditions such as menstruation and pregnancy, which are related to peritonitis, can also cause increased levels of CA125 [8]. This leads to the biomarker being quite specific but lacking sensitivity. In 2024, Aruni Ghose and his team demonstrated that the combination of patient characteristics, CA125, and HE4 provides higher specificity and sensitivity in multivariate index measurements than using CA125 and HE4 on their own [9]. In 2009, Moore developed a novel computational method called ROMA (risk of ovarian malignancy algorithm), aimed at more accurately diagnosing ovarian cancer [10]. In 2014, a research team led by Kaijser conducted a comprehensive analysis aimed at evaluating the effectiveness of the ROMA tool in diagnosing ovarian cancer [11].

On 12 October 2023, scientists from various countries convened in Washington, D.C. to deliberate on the ongoing discourse regarding the prospective benefits of artificial intelligence in the healthcare sector [12]. From the development of protein biomarkers to clinical trials and cancer diagnosis, artificial intelligence is increasingly being used as a medical assistance tool [13]. AI algorithms are capable of processing immense quantities of medical data at an astounding rate, which results in more efficient examinations and a reduction in the time required for diagnosis [14]. The accuracy of the AI breast cancer screening system was found to surpass that of all radiologists in the study by McKinney et al. in a study that evaluated the system [15]. Another experiment demonstrated that the adenoma detection rate of expert endoscopists was enhanced by AI-assisted colonoscopy. Researchers are currently investigating the potential of this technology to enhance ovarian cancer screening methods, as it has demonstrated advantages in specific disease areas [16].

There are several challenges in traditional machine learning approaches for ovarian cancer diagnosis [17]. These challenges include the difficulty in achieving high predictive accuracy due to the complexity and heterogeneity of ovarian cancer data, leading to issues in accurately classifying patients or predicting disease progression. Additionally, identifying relevant features or biomarkers from high-dimensional and diverse datasets can be challenging using traditional machine learning methods, potentially resulting in suboptimal model performance. Furthermore, certain traditional machine learning models, such as deep learning approaches, are often considered "black box" models, making it challenging to interpret the underlying decision making process. This lack of interpretability can hinder the understanding of the factors driving the model's predictions, which is critical in clinical settings. Moreover, imbalanced datasets, where the number of samples in different classes is disproportionate, can affect the performance of traditional machine learning models, leading to biased predictions and reduced generalization to new data. Due to the frequent issue of data imbalance in ovarian cancer datasets, models may struggle to fully learn their features. An effective remedy to this issue has recently been demonstrated by generative adversarial networks (GANs), which present a new strategy for improving datasets [18]. Additionally, considering the high dimensionality of ovarian cancer gene data, simplifying the dataset by selecting key features can enhance the model's learning efficiency. However, the training process for GANs is not always smooth sailing, and there is currently a lack of clear standards to measure the quality of the generated data. To address these challenges, we proposed an integrated algorithm with feature selection, data augmentation, and XGBoost for ovarian cancer within a comprehensive framework to optimize diagnostic accuracy. In the proposed approach, it can effectively integrate data augmentation with feature selection. On the other hand, XGBoost, an optimized gradient boosting technique, has demonstrated efficiency and applicability in feature selection and classification tasks [19].

The following is the structure of the remaining portion of this document. The fundamental principles of GAN and XGBoost, as well as the essential information of the dataset, are introduced in Section 2. The proposed technique is elaborated upon in Section 3. The proposed model's performance on the dataset is described in Section 4, which also contrasts it with other methodologies. Section 5 comprises the conclusion. Lastly, Section 6 outlines prospective future research directions.

## 2. Material and Methods

The dataset used in this study was sourced from six ovarian cancer microarray datasets provided on the Kaggle website: GSE6008, GSE9891, GSE18520, GSE38666, GSE66957, and GSE69428 [20]. To reduce systematic bias in the data, the PyComBat tool for batch effect correction was used to process these datasets. To safeguard patient privacy, all data utilized in this investigation were anonymized. This dataset covers five subtypes of ovarian cancer, comprising a total of 502 case samples, with each sample containing 11,476 feature points. Although the Kaggle method also uses feature selection and data augmentation, it focuses more on local feature processing and selecting individual dimensions. In contrast, our approach combines feature selection, GAN, and the XGBoost classifier to propose a holistic integrated framework. This integrated approach not only improves the model's accuracy but also effectively addresses the issue of class imbalance, making the synthetic samples more diverse and reducing the over-reliance on the dominant class. At the same time, we study the impact of different dimensions on model prediction accuracy, adopting a more comprehensive perspective to explore how various dimensions influence the overall performance of the model.

To minimize noise and interference from irrelevant features, we used a highly effective feature selection method (random forest) to filter the features. Figure 1 shows the distribution of the raw ovarian cancer data, which explains the imbalance present in the dataset. The quantity of samples in the three categories is substantially greater than that of the other classes. Figure 2 depicts the Pearson correlation heatmap of features between various features in the ovarian cancer dataset, showing that the most highly correlated scoring features have a close connection with the target feature category.

It also visualizes the relationships between these features. Figure 3 illustrates the degree of separation between different categories in the raw ovarian cancer data, explaining that different types of ovarian cancer exhibit distinct characteristics. The top 10 factors selected by the RF model are presented in Table 1, along with a concise description of their functions in the human body, based on their importance ranking.

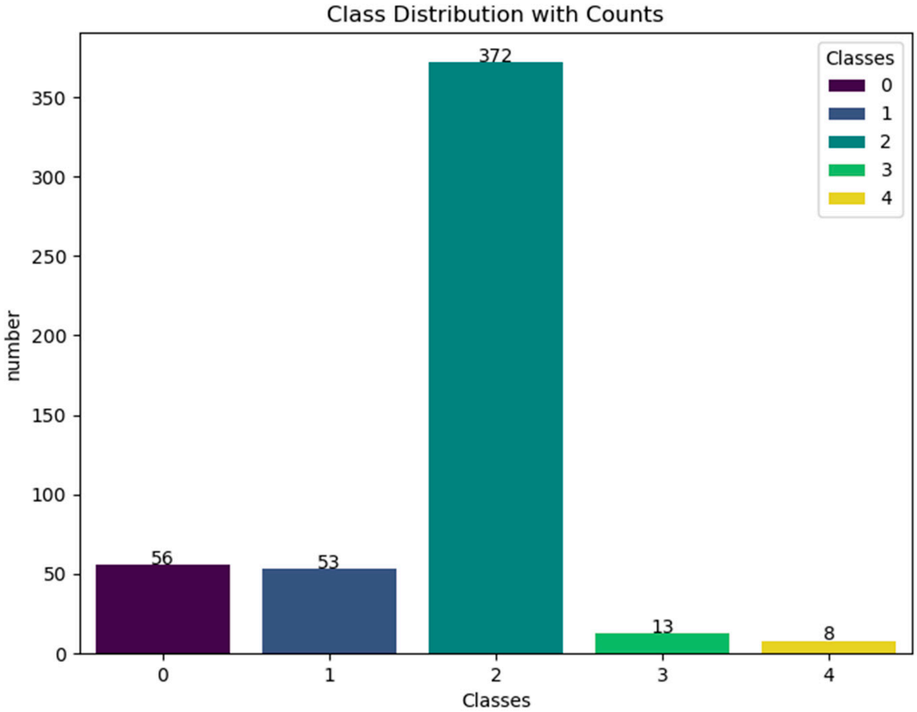

**Figure 1.** The distribution of ovarian cancer data.

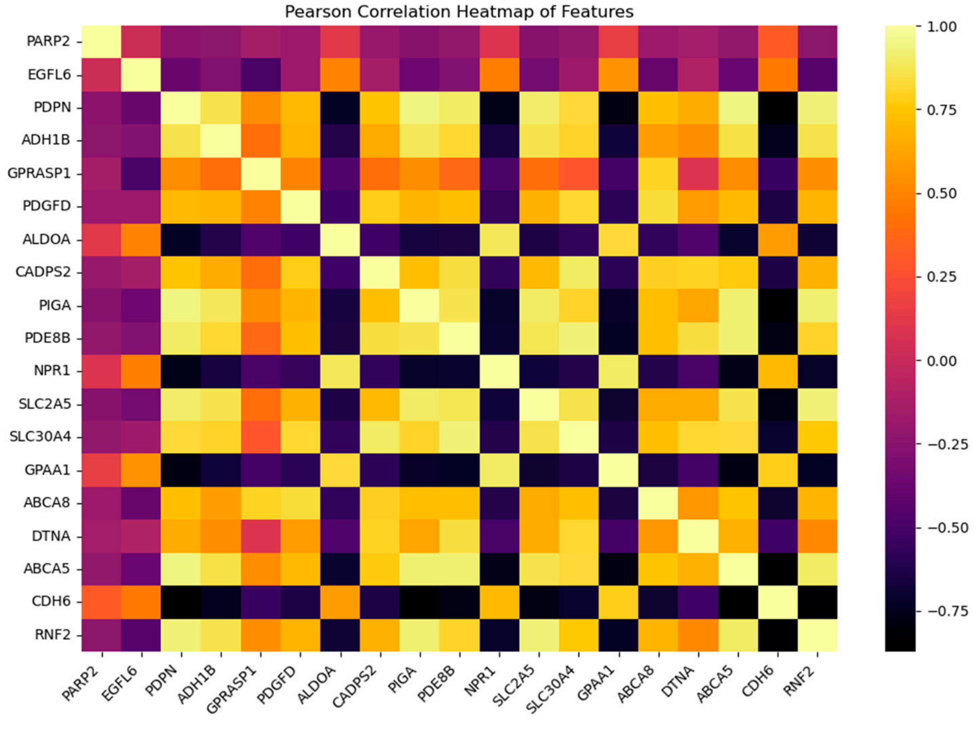

**Figure 2.** The Pearson correlation heatmap of features in ovarian cancer data.

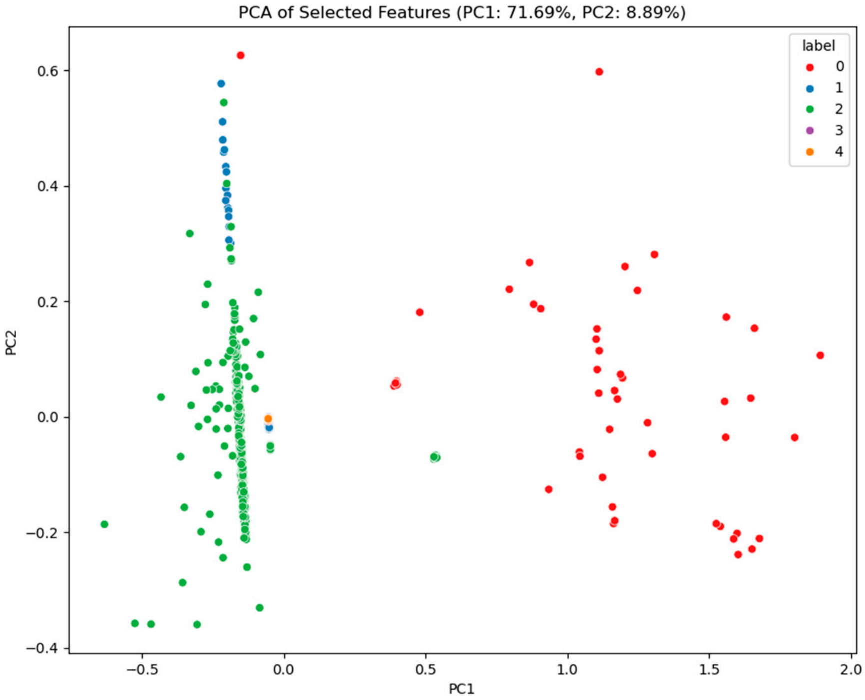

**Figure 3.** PCA distribution plot of the ovarian cancer data.

**Table 1.** The traits of ovarian cancer data.

| NUMBER | Feature Name | Note |
|---|---|---|
| 1 | FTCD | Helps with folate metabolism, important for DNA synthesis. |
| 2 | KRTAP5-8 | Related to hair and skin structure. |
| 3 | GSK3A | Regulates cell metabolism and growth. |
| 4 | RAB5A | Involved in cell transport and endocytosis. |
| 5 | SNAPC5 | Important for gene transcription. |
| 6 | CHTOP | Regulates RNA and gene expression. |
| 7 | GPR137 | A receptor that may affect immune responses. |
| 8 | GNAQ | As part of signaling pathways; mutations are found in some cancers. |
| 9 | UBAP1 | Involved in protein tagging and transport in cells. |
| 10 | IGHMBP2 | Helps with RNA processing; linked to neurological disorders. |

Since this paper is based on data augmentation using ACGAN and classification using XGBoost, this section provides a brief introduction to these two components.

### 2.1. Description of XGBoost in Brief

XGBoost (extreme gradient boosting), proposed by Tianqi Chen in 2016, is an efficient machine learning algorithm designed to improve the gradient boosting performance in distributed computing and sparse data handling [21]. At its core, XGBoost is based on gradient boosting decision trees (GBDT), which utilize both the gradient and the second-order derivative (Hessian) from Taylor expansion. This dual information enhances optimization accuracy, accelerates convergence, and aids in split gain calculation. The objective function of XGBoost includes both the loss function and regularization, expressed as follows:

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{t=1}^{T} \Omega(f_t) \tag{1}$$

$$l\left(y_i, \hat{y}_i\right) = log\left(1 + exp\left(-y_i \hat{y}_i\right)\right) \tag{2}$$

Here, $n$ represents the number of samples, $y_i$ is the true label of the $i$-th sample, and $\hat{y}_i$ is the predicted value for the $i$-th sample. $l$ denotes different loss functions; in this experiment, we use the logarithmic loss function. $T$ represents the total number of trees, and *it* is the prediction function of the $t$-th tree. The term $\Omega$ indicates the regularization term, which is specified as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2 \tag{3}$$

In this context, $\gamma$ represents the regularization coefficient for the complexity of the tree, $\lambda$ is the L2 regularization coefficient for the weights of the leaf nodes, and $\omega$ is the weight of the leaf nodes. XGBoost, during the process of adding tree models, first calculates the negative gradient based on the residual of each sample, and the new tree fits this negative gradient to minimize the loss function. The construction of the tree structure is carried out using a greedy algorithm, where at each split, the algorithm evaluates all possible split points and features, selecting the split point that most reduces the objective function, and finally, calculates the predictive value of the model. The iterative process involves adding the newly constructed tree model to the existing model and updating the model's predictive value. For the $i$-th example, $f_t(x_i)$ shows how well the new tree model can predict what will happen. The formula is as follows:

$$\hat{y}_i \leftarrow \hat{y}_i + f_t(x_i) \tag{4}$$

The final prediction of the entire model is the weighted sum of the outputs of all the trees, which is represented as follows:

$$\hat{y}_i = \phi(x_i) = \sum_{t=1}^{T} f_t(x_i) \tag{5}$$

### *2.2. Description of GAN in Brief*

Generative adversarial networks (GANs), introduced by Ian Goodfellow and colleagues [22], are powerful generative models. The core idea behind GANs is a dynamic game between two key players: the generator (G) and the discriminator (D). These neural networks engage in adversarial training, learning to reach a Nash equilibrium. The objective function of GANs is a minimax problem, representing an adversarial process between G and D, as shown in the following formula:

$$\underset{G}{min}\underset{D}{max} V(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1-D(G(z)))] \tag{6}$$

In this context, $x$ represents a sample from the real data, and $z$ is a random noise vector drawn from the latent space. $p_{data}(x)$ is the distribution of the real data. $p_z(z)$ is the distribution of the random noise, which is typically assumed to be a normal distribution. D($x$) is the discriminator's judgment (probability) that the real sample $x$ comes from the real data. D(G($z$)) is the discriminator's judgment that the generated sample G($z$) comes from the real data. The minimax objective of the entire training process means that, for the discriminator network, its goal is to maximize its correct judgment rate, that is, to be able to identify real images and images generated by the generator. The generator labeled as G in a GAN plays the role of generating data, with its core objective being to produce data that closely resembles the original samples, making it difficult for the discriminator to distinguish between real and fake data. Figure 4 illustrates the GAN's particular architecture.

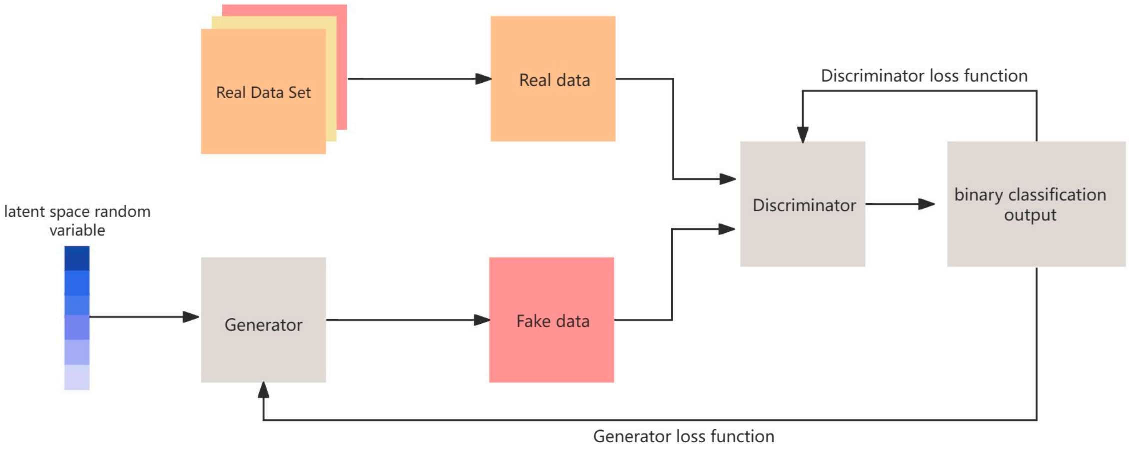
Real Data Set
Real data
Discriminator loss function
Discriminator
binary classification output
latent space random variable
Generator
Fake data
Generator loss function

**Figure 4.** The basic architecture of GAN.

The original GAN has limited capabilities, generating only simple, single data based on input. Koo et al. used deep convolutional GANs (DCGANs) to colorize black-and-white images [23]. The introduction of conditional GAN (cGAN) allowed for the generation of samples that not only match the training data distribution but also meet specific conditions [24]. In 2016, Odena and colleagues proposed the ACGAN structure, which added an auxiliary classifier, enabling GAN to generate samples that align with the training data distribution and fulfill specific conditions [25]. Figure 5 illustrates the ACGAN architecture. By providing both the generator and discriminator with conditional information, ACGAN improves the quality and diversity of generated data. During discriminator training, an additional classification loss is used alongside the binary cross-entropy loss, helping the model learn richer feature representations. In this study, ACGAN is employed for data augmentation, generating diverse ovarian cancer data to improve the accuracy of classifiers in distinguishing between cancer types.

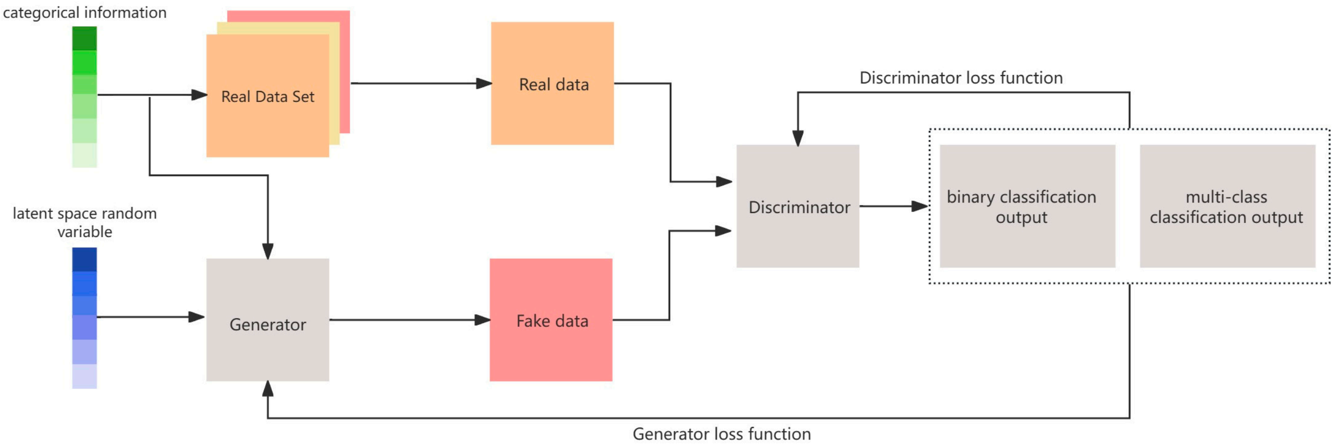
categorical information
Real Data Set
Real data
Discriminator loss function
Discriminator
binary classification output
multi-class classification output
latent space random variable
Generator
Fake data
Generator loss function

**Figure 5.** The basic architecture of ACGAN.

## 3. The Proposed Algorithm

The ovarian cancer dataset studied in this research encompasses five distinct types of ovarian cancer and over 10,000 associated genes. If we were to apply this to real-world scenarios, the sheer volume of genetic data would make it extremely challenging for physicians to provide timely preventive diagnoses for their patients. Moreover, as shown in Figure 1, the number of cases of various types in the sample varies greatly. This imbalance can result in the model performing inadequately in learning to recognize those ovarian cancer types that are under-represented in terms of sample size. To address these challenges, we employ feature selection and data augmentation techniques to process the dataset.

Feature selection lowers data complexity, augmenting model interpretability, while GAN data augmentation amplifies sample quantity without modifying intrinsic features, thereby strengthening the model's capacity to learn from underrepresented classes. The processed dataset is then classified using the XGBoost model, a highly efficient machine learning algorithm known for its effectiveness in dealing with imbalanced datasets. Figure 6 illustrates the flowchart of the entire algorithm, detailing each step from data preprocessing to model training and evaluation. By using this method, we aim to boost the model's precision in distinguishing many forms of ovarian cancer, therefore offering more consistent support for clinical diagnosis.

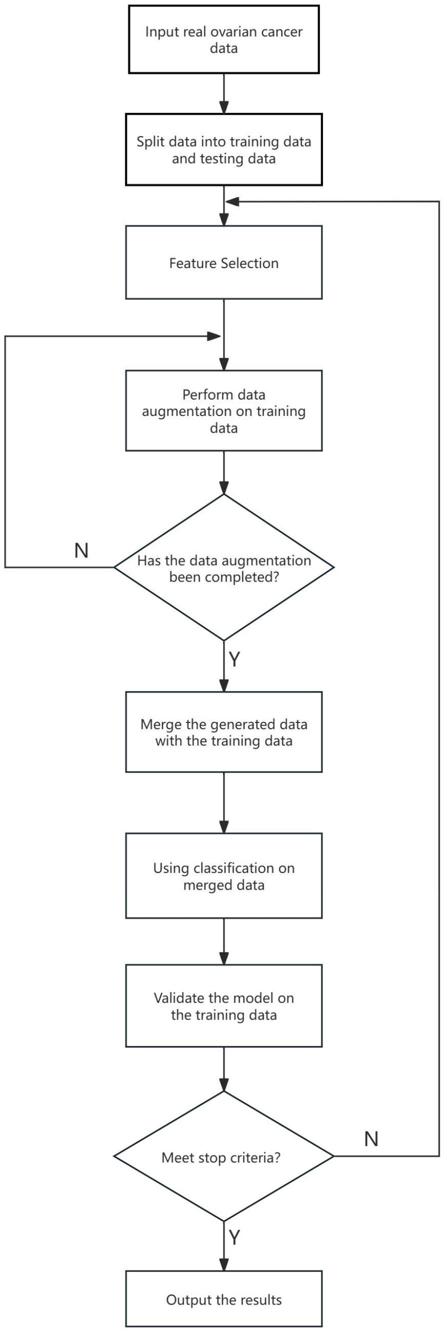
Input real ovarian cancer data
Split data into training data and testing data
Feature Selection
Perform data augmentation on training data
Has the data augmentation been completed?
N
Y
Merge the generated data with the training data
Using classification on merged data
Validate the model on the training data
Meet stop criteria?
N
Y
Output the results

**Figure 6.** The flowchart of the integrated algorithm proposed in this paper.

We begin in Figure 6 by importing the raw ovarian cancer data and then splitting the dataset into training and validation sets before performing feature selection. As is well known, in high-dimensional datasets, many features may be noise. Selecting features with high relevance can effectively reduce the risk of model overfitting, and feature selection can reduce the computational resources required for model training and prediction, as well as reduce the workload of doctors in real-life situations.

Currently, there are three mainstream feature selection methods: filtering methods, wrapper methods, and embedded methods. When it comes to these approaches, embedded methods are the ones that carry out feature selection within the process of model training, as opposed to doing it as a separate pre-processing step. This approach allows for the consideration of the relationship between features and the model, as well as the interaction between features, thereby selecting the features that are most helpful to the model's performance. We performed comparative studies to investigate in this approach the effect of the dimensionality after feature selection on model correctness. Considering the workload of doctors during diagnosis in real life, we chose to limit the number of factors to no more than fourteen. Figure 7 shows the effect of different dimensions of feature selection on the model's classification accuracy. It can be observed that when the dimension is twelve, the model's accuracy rate reaches about 99%, and then, it decreases at thirteen dimensions. Therefore, in this study, we used the RF method within embedded methods to screen the dataset, resulting in a dataset with eight feature factors. During the training process, random forest evaluated the importance of each feature based on its contribution to the data splits. For each feature, random forest accumulated the gains or reductions in impurity during the training, resulting in an importance score. After training, RF computed an importance score for each feature, which reflected its contribution to the model's prediction. The importance was calculated by measuring the Gini index for each feature in the decision trees. Figure 8 is the RF model's scoring and ranking of all factors in the dataset, selecting the top eight feature factors from high to low. It can be seen that FTCD is the most important factor.

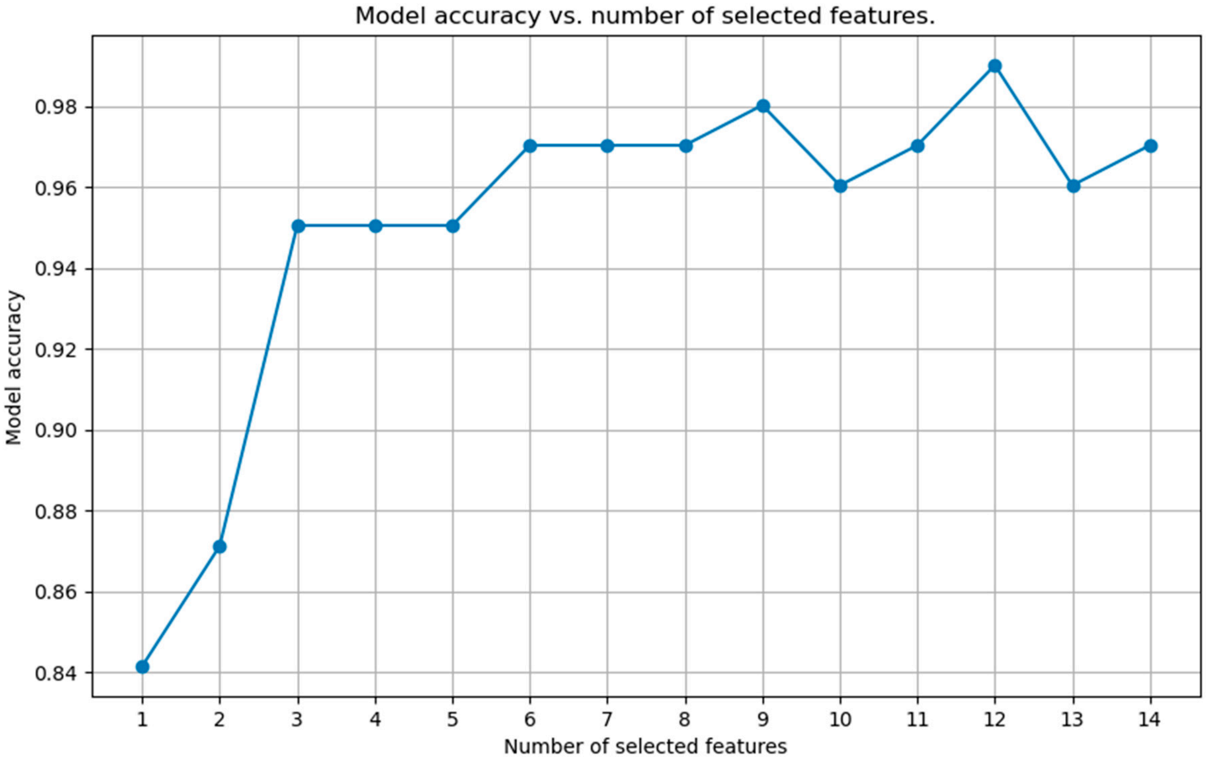

**Figure 7.** Model accuracy vs. number of selected features.

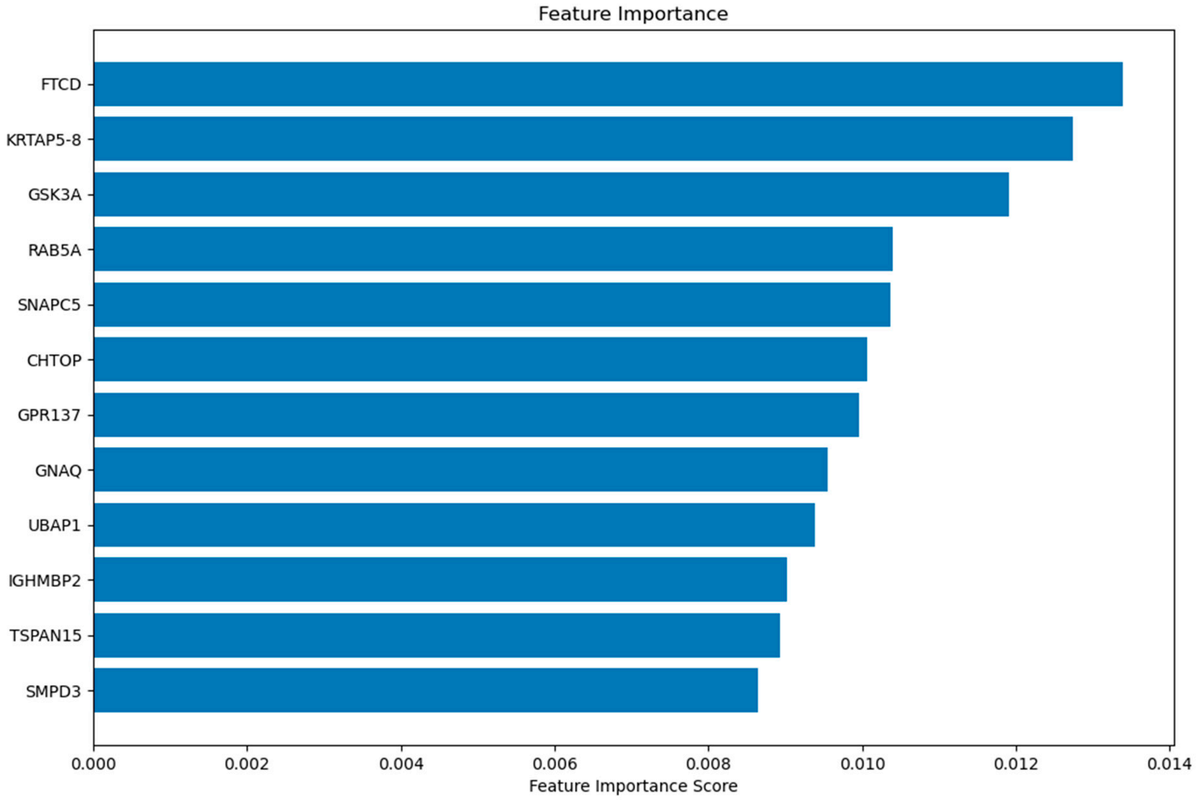

**Figure 8.** Features after feature selection.

There are various methods of data augmentation, including SMOTE, etc., and the core idea behind these methods is linear interpolation. Such methods generate relatively simple samples, lack diversity, and are prone to creating unrealistic or repetitive samples, which can affect subsequent classification predictions. Our data augmentation was carried out by training ACGAN on the original data to specify the GAN network to generate a corresponding number of ovarian cancer type samples. ACGAN introduced an auxiliary classifier to help the discriminator better learn the categories of the generated data, which can fully address the issue of data imbalance and improve the model's training for all types. In Table 2, we can see that, under the same feature selection and classifier environment, using ACGAN as a data augmentation technique is more effective than using SMOTE.

**Table 2.** Comparison of different data augmentation techniques.

| Method | Classification Accuracy |
|---|---|
| SMOTE | 98.02% |
| ACGAN | 99.01% |

We integrated the data produced by the GAN network with the training set derived from the original data to create a new dataset that eliminates data imbalance, subsequently inputting it into the XGBoost classification model, a robust machine learning algorithm renowned for its exceptional performance in classification tasks. Overall, the prediction steps for ovarian cancer by the entire model can be divided into the following steps:

Step 1: Import the original dataset, process it in batches via PyComBat, then partition the dataset into 80% for training and 20% for validation.

Step 2: Import the processed dataset into the RF model and select eight important factors based on feature scores.

Step 3: Import the training set after feature selection into the ACGAN model, specify the generation of 1000 samples for each type, then merge the generated data with the original training set, and input it into the XGBoost model for training.

Step 4: Use the trained model to classify the validation set and evaluate the results.

## 4. Results and Discussions

The experiment utilized an AMD Ryzen 7 8-core processor clocked at 4.75 GHz, alongside an NVIDIA GTX 4060 GPU with dedicated video memory. The version numbers utilized in this study have been provided, encompassing PyComBat 0.20, Keras 2.7.0, Python 3.9.18, Numpy 1.26.3, Pandas 2.1.4, XGBoost 2.1.0, and TensorFlow-GPU 2.7.0. It is also worth noting that the SMOTE technique was incorporated using the imblearn library, version 0.11.0. The ACGAN network model underwent 3000 epochs, using the Adam optimizer to adjust the weight parameters in the GAN model, with a learning rate set at 0.0002. Two varieties of cross-entropy loss functions were employed: binary cross-entropy, for distinguishing between authentic and counterfeit data, and categorical cross-entropy, for managing the loss function associated with sample categorization. We specified the GAN network to generate 1000 samples for each type of ovarian cancer, with the specific number of samples shown in Figure 9. After the original ovarian cancer data were augmented, the distribution of the sample data became more uniform. In Table 3, we list some key hyperparameters used in the models of this experiment.

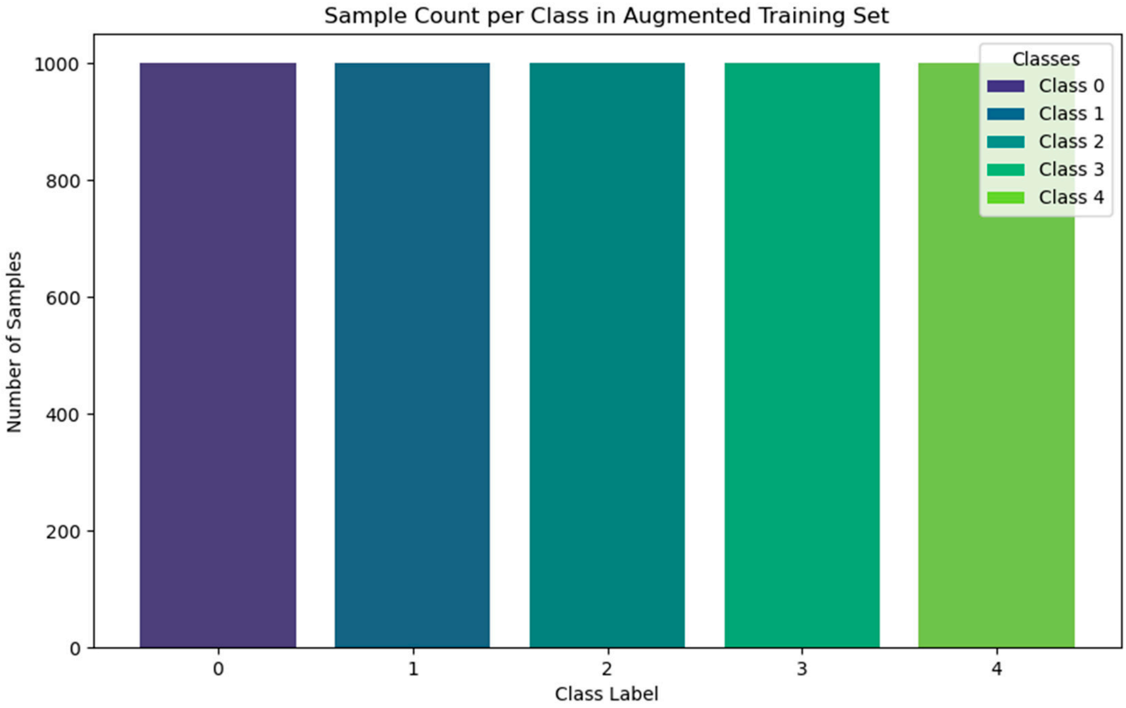

**Figure 9.** Distribution of ovarian cancer samples after ACGAN model.

To further verify the effectiveness of our model, we employed an m-fold cross-validation approach, with m = 5 to evaluate the performance during the experiment. The original dataset we imported was divided into five subsets. One subset was put aside as the test set in every cycle; the other four subsets were used as the training set to equip the model for subsequent validation on the test set. This procedure was executed five times, with a distinct subset as the test set, while the other four subsets functioned as the training set in each iteration. Table 4 shows the outcome of the experiment. In the proposed algorithm, the best-performing fold achieved perfection in accuracy (Acc), precision (Pre), recall (Rec), and F1 score. Furthermore, the variation in each fold was not substantial, demonstrating that the algorithmic model we proposed has strong classification capabilities and can accurately predict ovarian cancer diseases based on the factors. This consistency across folds indicates that our model is robust and reliable, as it performs well even when different subsets of the data are held out for testing. The model's high scores in accuracy, precision, and recall indicate that it performs excellently in detecting the presence or absence of ovarian cancer.

**Table 3.** Key hyperparameters of each model.

| Model | Key Hyperparameters | Description |
|---|---|---|
| Random Forest (Feature Selection) | n_estimators (100) | Number of estimators for feature selection |
| Discriminator Model | Dense layers: [64, 32], activation functions: [relu, sigmoid, softmax] | Discriminator model architecture with two outputs: fake/real (sigmoid) and class predictions (softmax) |
| Generator Model | Dense layers: [64, n_features], activation functions: [relu, linear] | Generator model with random input and class embedding |
| GAN Model | Learning rate (0.0002), beta_1 (0.5) | Learning rate and beta_1 for the Adam optimizer |
| XGBoost Classifier | n_estimators (100), random_state (42) | Number of estimators for the final classifier |

**Table 4.** The model's cross-validation results for the ovarian cancer dataset.

| Experiment | Acc % | Pre % | Rec % | F1 Score % |
|---|---|---|---|---|
| m-fold #1 | 98.98 | 98.99 | 98.98 | 98.96 |
| m-fold #2 | 100 | 100 | 100 | 100 |
| m-fold #3 | 100 | 100 | 100 | 100 |
| m-fold #4 | 100 | 100 | 100 | 100 |
| m-fold #5 | 99.90 | 99.24 | 99.09 | 99.07 |

To effectively illustrate the efficacy of the proposed algorithm, we will compare it with several prevalent categorization algorithms including those available on the Kaggle platform. To ensure the fairness of the comparative experiment, all algorithmic models use the same fold of the dataset and are trained and tested under the same experimental conditions. Table 5 shows the comparisons' outcomes. The approach on the Kaggle website also includes feature selection and data augmentation, similar to our methodology.

**Table 5.** The performance of different models.

| Method | Classification Accuracy |
|---|---|
| SVR | 68.32% |
| SVM | 78.22% |
| LR | 89.11% |
| Kaggle [26] | 94.06% |
| XGBoost | 95.05% |
| The proposed algorithm | **99.01%** |

From Table 5, we can conclude that the accuracy of the algorithm we proposed is superior to several other algorithms, including support vector regression (SVR), support vector machine (SVM), logistic regression (LR), XGBoost, and methods found on the Kaggle website. Support vector regression (SVR) is a technique used in the field of machine learning for regression analysis. The core idea of SVR is to allow a certain range of error between the predicted values and the actual values; errors within this range are not penalized. This approach helps to reduce sensitivity to noise and can improve the model's generalization ability [27]. Applied in both classification and regression, support vector machine (SVM) is a supervised learning technique. It seeks a hyperplane separating data points of many kinds, thereby guaranteeing the generalization capacity of the model and the accuracy of classification. SVM is known for its strong resistance to overfitting [28]. Logistic regression (LR) operates by applying a logistic function to the output of linear regression to determine the likelihood of a sample belonging to a certain class. Although LR is effective for binary or multiclass classification, its model structure is relatively simple [29]. It is worth noting that, although XGBoost is a highly effective classification model, the raw ovarian cancer data have a large number of features and exhibit significant data imbalance. If the raw data are fed directly into XGBoost for prediction without undergoing feature selection and data augmentation, the results are not very satisfactory. Kaggle is a renowned online platform for data science competitions, and there are algorithms on the site that process this dataset. However, the method they choose involves feature selection first, narrowing down to seven factors and then performing hyper-parameter tuning on the XGBoost model using Optuna. The results show that this approach is not very ideal. As can be seen from Figure 3, different algorithm models exhibit varying classification accuracy rates for the same data. The classification accuracy rates of SVR, SVM, and LR are 68.32%, 78.22%, and 89.11%, respectively. Since these algorithms are more traditional machine learning algorithms, their classification effects are not satisfactory. In the latest research, our proposed algorithm has achieved remarkable results in the accuracy of ovarian cancer detection, with an accuracy rate of 99.01%. This figure is significantly higher than the 95.05% reported on the Kaggle website and the 94.06% achieved by the XGBoost algorithm. These results not only highlight the significant advantages of our developed algorithm in terms of accuracy compared to other methods but also imply that the algorithm may play a key role in the early identification and intervention of ovarian cancer in the future, potentially having a positive impact on improving treatment outcomes and patient survival rates.

To further assess the performance of our proposed algorithm on the ovarian cancer dataset, we have constructed a confusion matrix and a receiver operating characteristic (ROC) curve. The confusion matrix, depicted in Figure 10, is a 5 × 5 grid, reflecting the five distinct classes of ovarian cancer within the dataset. The model's accurately predicted samples are represented by the entries along the diagonal of the matrix. Upon examining Figure 10, it is evident that the samples for classes 0, 1, 2, and 3 are accurately predicted, except for one sample from class 3, which is mistakenly identified as class 4.

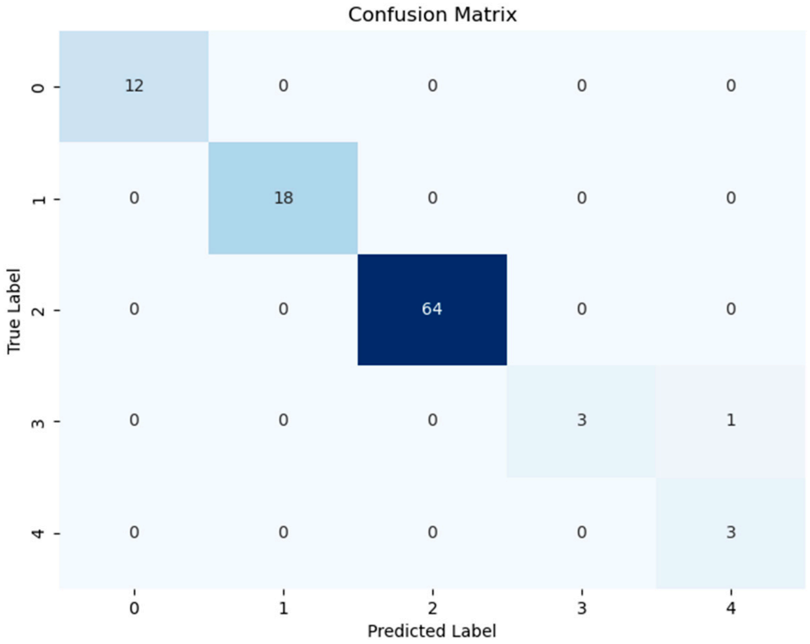

**Figure 10.** The confusion matrix.

The ROC curve in Figure 11 shows the model's classification performance at different threshold settings. The *x*-axis represents the false positive rate (FPR), while the *y*-axis represents the true positive rate (TPR). As the decision threshold varies, the TPR and FPR change, creating the curve. The area under the curve (AUC) is a metric of the model's ability to discriminate between classes, with values ranging from 0 to 1. A higher AUC indicates better classification ability, while an AUC of 0.5 suggests no discriminative power.

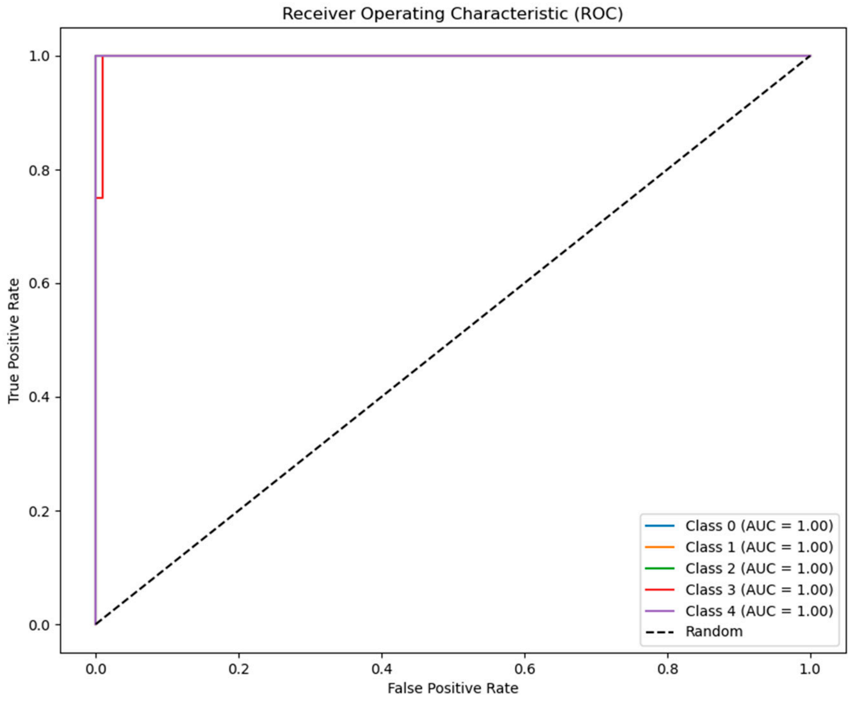

**Figure 11.** The ROC curve.

The ROC curve in Figure 11 shows that the AUC values for all categories have reached a perfect score of 1, indicating that the classifiers for these categories are very ideal, with only the ROC curve for category 3 slightly deviating. However, the AUC for category 3 still shows a value of 1, which is due to the small number of cases in category 3, leading to a potentially inflated AUC during calculation. These findings, as shown in Figure 11, confirm the strong predictive capability of the algorithm developed in this study for classifying ovarian cancer types.

To validate the model's generalizability and to prevent the synthetic samples from excessively reflecting the characteristics of the dominant class, we also evaluated the model using a balanced dataset. This dataset, also from Kaggle, is the Wisconsin Breast Cancer dataset, which contains 569 cases, with 212 malignant (cancerous) cases and 357 benign (non-cancerous) cases [30]. To ensure class balance, we randomly reduced the number of benign cases to match the number of malignant cases. The processed dataset was then fed into the model, and as shown in Table 6, the model achieved an accuracy of 97.73%, which is still higher than the accuracy of the model on Kaggle, ensuring the robustness of the proposed model.

**Table 6.** The performance of the proposed models in breast cancer.

| Method | Classification Accuracy |
|---|---|
| Kaggle [30] | 97.36% |
| The proposed algorithm | **97.73%** |

## 5. Conclusions

Ovarian cancer is a prevalent gynecological malignancy globally, distinguished by its significantly high mortality rate. The disease is often challenging to detect early on, because its initial symptoms are rarely noticeable, which results in greater treatment difficulties and substantially increased medical expenses in later stages. This situation places a tremendous financial and emotional burden on patients and their families. However, by identifying early-stage ovarian cancer through prediction and screening, we can not only significantly enhance patients' survival rates and quality of life but also provide doctors with sufficient time to devise more tailored treatment plans. The algorithmic model proposed in this paper is capable of efficiently predicting ovarian cancer based on genetic factors. It integrates feature selection, data augmentation, and the XGBoost module, enabling accurate forecasting of the specific types of ovarian cancer. This model boasts a higher classification accuracy rate compared to other mainstream algorithms, achieving a remarkable classification accuracy rate of 99.01%. This sophisticated model is a viable instrument for the early identification of ovarian cancer, possibly revolutionizing treatment approaches and enhancing patient outcomes. By leveraging cutting-edge machine learning techniques, it empowers medical professionals with the ability to identify cancer at its most treatable stages, thereby saving lives and reducing the overall impact of this deadly disease.

By introducing feature selection to reduce the dimension of the original dataset, we can decrease the consumption of feature extraction and simultaneously suppress the interference of useless features on the model. Through feature selection, we can identify the top three factors in the dataset as FTCD, KRTAP-8, and GSK3A. Mostly connected with the metabolism of folic acid in the human body, FTCD is a multifunctional enzyme that catalyzes two successive events in the one-carbon metabolism pathway. Changes in its expression may be related to the occurrence of ovarian cancer, as its connection to the folic acid metabolic pathway makes it an important target for studying tumor metabolism. As a member of the keratin-associated protein family, KRTAP-8 may influence cancer cell adhesion and migratory capacity, given changes in the tumor microenvironment and cellular structure. The protein encoded by the GSK3A gene is a serine and threonine kinase that plays a central role in regulating cell signaling, protein synthesis, and cytoskeletal dynamics. Although GSK3A plays multiple roles within the cell, it is particularly noted in cancer research for its potential tumor-suppressive properties. All things considered, the elements chosen by feature screening may be used as markers for early ovarian cancer detection in patients, thereby significantly lowering the influence of irrelevant elements and improving the classification capacity of the model.

Our algorithm aims to predict ovarian cancer through patients' genetic data, representing a cutting-edge approach based on biomarkers for early detection and personalized treatment. The development of artificial intelligence technology has provided new hope for the early detection and diagnosis of ovarian cancer, but issues of data quality and bias remain obstacles to the future implementation of AI-assisted diagnosis [31]. Such issues are particularly problematic in the medical field, as they can lead to diagnostic disparities and selection biases among different populations [32]. Consequently, while the algorithm has attained an accuracy rate of 99.01% in predicting ovarian cancer, our objective is not to supplant the clinical judgment of medical professionals but to alleviate their workload in managing extensive genetic data, thereby offering them supplementary information and insights. Utilizing the algorithm presented in this research to analyze the ovarian cancer gene dataset enables physicians to accurately detect ovarian cancer in its first stages. Finding the genetic causes of ovarian cancer would enable physicians to create novel therapeutic medications and specifically targeted treatment approaches, therefore improving the survival rate and quality of life for each patient.

In summary, the ovarian cancer prediction model developed in this study holds significant clinical importance for improving early diagnosis and personalized treatment. Although this model has demonstrated extremely high accuracy compared to other processing methods, due to the limitations of the dataset, we still need to further verify and optimize it in a broader patient population. It is believed that, with the continuous advancement of technology and in-depth research, this gene-based prediction model will be able to reduce people's fear of ovarian cancer and bring them greater hope.

## 6. Future Research Directions

In future research, we aim to enhance the generalization of our models through multicenter data collection across different regions and ethnicities, ensuring a diverse and extensive dataset. We will implement rigorous standardization processes to ensure the model's robust application in various clinical environments. As AI continues to integrate into healthcare, greater attention must be paid to regulatory frameworks and safeguarding data privacy and security [33]. We plan to adopt cutting-edge encryption and anonymization techniques to protect patient privacy and ensure compliance with international data protection regulations. Additionally, we will prioritize collecting data from rare ovarian cancer cases and across various stages of the disease, while also considering the influence of environmental and lifestyle factors on disease progression. These efforts will provide richer training data, improving the model's accuracy and reliability for clinical use.

**Author Contributions:** Conceptualization, J.C.; Methodology, Z.-J.L.; Formal analysis, Z.L.; Investigation, J.C. and Y.L.; Resources, Y.L.; Data curation, C.-H.H.; Writing—review & editing, J.C and Z.-J.L.; Supervision, Z.-J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by Fuzhou University and partly supported by the Fujian Province Natural Science Foundation of China under grant no. 2024J08236.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are available at https://www.kaggle.com/code/yoshifumimiya/multiclass-imbalanced-data-vol-1 (accessed on 10 October 2024) and https://www.kaggle.com/code/rahmasleam/breast-cancer-prediction-using-automl-models (accessed on 18 December 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2024**, *74*, 229–263. [CrossRef] [PubMed]
2. Torre, L.A.; Trabert, B.; DeSantis, C.E.; Miller, K.D.; Samimi, G.; Runowicz, C.D.; Gaudet, M.M.; Jemal, A.; Siegel, R.L. Ovarian cancer statistics, 2018. *CA Cancer J. Clin.* **2018**, *68*, 284–296. [CrossRef] [PubMed]
3. Jayson, G.C.; Kohn, E.C.; Kitchener, H.C.; Ledermann, J.A. Ovarian cancer. *Lancet* **2014**, *384*, 1376–1388. [CrossRef] [PubMed]
4. Dochez, V.; Caillon, H.; Vaucel, E.; Dimet, J.; Winer, N.; Ducarme, G. Biomarkers and algorithms for the diagnosis of ovarian can-cer: CA125, HE4, RMI and ROMA, a review. *J. Ovarian Res.* **2019**, *12*, 28. [CrossRef] [PubMed]
5. Grossman, D.C.; Curry, S.J.; Owens, D.K.; Barry, M.J.; Davidson, K.W.; Doubeni, C.A.; Epling, J.W., Jr.; Kemper, A.R.; Krist, A.H.; Kurth, A.E.; et al. Screening for Ovarian Cancer: US Preven-tive Services Task Force Recommendation Statement. *JAMA* **2018**, *319*, 588–594.
6. Rosenthal, A.N.; Fraser, L.S.; Philpott, S.; Manchanda, R.; Burnell, M.; Badman, P.; Hadwin, R.; Rizzuto, I.; Benjamin, E.; Singh, N.; et al. Evidence of Stage Shift in Women Diagnosed with Ovarian Cancer During Phase II of the United Kingdom Familial Ovarian Cancer Screening Study. *J. Clin. Oncol.* **2017**, *35*, 1411–1420. [CrossRef]
7. Bast, R.C.; Feeney, M.; Lazarus, H.; Nadler, L.M.; Colvin, R.B.; Knapp, R.C. Reactivity of a monoclonal antibody with human ovarian carcinoma. *J. Clin. Investig.* **1981**, *68*, 1331–1337. [CrossRef]
8. Buamah, P. Benign conditions associated with raised serum CA-125 concentration. *J. Surg. Oncol.* **2000**, *75*, 264–265. [CrossRef]
9. Ghose, A.; McCann, L.; Makker, S.; Mukherjee, U.; Gullapalli, S.V.N.; Erekkath, J.; Shih, S.; Mahajan, I.; Sanchez, E.; Uccello, M.; et al. Diagnostic biomarkers in ovarian cancer: Advances beyond CA125 and HE4. *Ther. Adv. Med. Oncol.* **2024**, *16*, 17588359241233225. [CrossRef]
10. Moore, R.G.; McMeekin, D.S.; Brown, A.K.; DiSilvestro, P.; Miller, M.C.; Allard, W.J.; Gajewski, W.; Kurman, R.; Bast, R.C., Jr.; Skates, S.J. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol. Oncol.* **2009**, *112*, 40–46. [CrossRef]
11. Kaijser, J.; Van Belle, V.; Van Gorp, T.; Sayasneh, A.; Vergote, I.; Bourne, T.; Van Calster, B.; Timmerman, D. Prognostic value of serum HE4 levels and risk of ovarian malignancy algorithm scores at the time of ovarian cancer diagnosis. *Int. J. Gynecol. Cancer* **2014**, *24*, 1173–1180. [CrossRef] [PubMed]
12. AI for Scientific Discovery—Oct. 12–13 Workshop. National Academies. Available online: https://www.nationalacademies.org/news/2023/10/ai-for-scientific-discovery-oct-12-13-workshop (accessed on 30 October 2023).
13. Amisha; Malik, P.; Pathania, M.; Rathaur, V.K. Overview of artificial intelligence in medicine. *J. Fam. Med. Prim. Care* **2019**, *8*, 2328–2331. [CrossRef] [PubMed]
14. Huang, S.; Yang, J.; Fong, S.; Zhao, Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.* **2020**, *471*, 61–71. [CrossRef] [PubMed]

15. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef]
16. Xu, H.; Tang, R.S.Y.; Lam, T.Y.T.; Zhao, G.; Lau, J.Y.W.; Liu, Y.; Wu, Q.; Rong, L.; Xu, W.; Li, X.; et al. Artificial intelligence–assisted colonoscopy for colorectal cancer screening: A multicenter randomized con-trolled trial. *Clin. Gastroenterol. Hepatol.* **2023**, *21*, 337–346.e3. [CrossRef]
17. Feng, Y. An integrated machine learning-based model for joint diagnosis of ovarian cancer with multiple test indicators. *J. Ovarian Res.* **2024**, *17*, 45. [CrossRef]
18. Sun, Z.; Wang, G.; Li, P.; Wang, H.; Zhang, M.; Liang, X. An improved random forest based on the classification accuracy and cor-relation measurement of decision trees. *Expert. Syst. Appl.* **2024**, *237*, 121549. [CrossRef]
19. Asselman, A.; Khaldi, M.; Aammou, S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact. Learn. Environ.* **2023**, *31*, 3360–3379. [CrossRef]
20. Ovarian Cancer Datasets. Available online: https://www.kaggle.com/datasets/yoshifumimiya/6-ovarian-cancer-datasets/data (accessed on 10 October 2024).
21. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
23. Koo, S. Automatic colorization with deep convolutional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 212–217.
24. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784, 2014.
25. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis With Auxiliary Classifier GANs. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017.
26. Multiclass Imbalanced Data Vol.1. Available online: https://www.kaggle.com/code/yoshifumimiya/multiclass-imbalanced-data-vol-1 (accessed on 10 October 2024).
27. Jiang, P.; Ge, N. Evaluation of Fracture Energy of Concrete By Hybrid SVR Analysis. *J. Appl. Sci. Eng.* **2024**, *27*, 3645–3661.
28. Kollem, S. An efficient method for MRI brain tumor tissue segmentation and classification using an optimized support vec-tor machine. In *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2024.
29. Gao, J.; Gong, Z. Uncertain logistic regression models. *AIMS Math.* **2024**, *9*, 10478–10493. [CrossRef]
30. Breast Cancer Prediction Using AutoML Models. Available online: https://www.kaggle.com/code/rahmasleam/breast-cancer-prediction-using-automl-models (accessed on 18 December 2024).
31. Gianfrancesco, M.A.; Tamang, S.; Yazdany, J.; Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **2018**, *178*, 1544–1547. [CrossRef] [PubMed]
32. Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; Tsaneva-Atanasova, K. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **2019**, *28*, 231–237. [CrossRef]
33. Murdoch, B. Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Med. Ethic* **2021**, *22*, 122. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.