IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 10, 2021, accepted August 22, 2021, date of publication August 24, 2021, date of current version September 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3107498

A Machine Learning-Aided Framework to Predict Outcomes of Anti-PD-1 Therapy for Patients With Gynecological Cancer on Incomplete Post-Marketing Surveillance Dataset

XIAOMEI LIU[1], ZHIFENG XIAO[2], YANG SONG[1], RUIZHE ZHANG[1], XIUQIN LI[1], AND ZHENHUA DU[1]

[1]Department of Obstetrics and Gynecology, Sheng Jing Hospital, China Medical University, Shenyang, Liaoning 110004, China
[2]School of Engineering, Penn State Erie, The Behrend College, Erie, PA 16563, USA

Corresponding authors: Xiuqin Li (2403621353@qq.com) and Zhenhua Du (duzhenhua0000@163.com)

This work was supported by the Liaoning Science and Technology Department, China, through the project titled "Research on the Mechanism of VH4-34 Encoding Human IgM Anti-Cancer Immune Micro-Environment," under Grant 2020-BS-099 and Grant 2020.05.01-2023.04.30.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB).

ABSTRACT Post-marketing surveillance of antineoplastic agents is performed to evaluate the efficacy and safety in patients aiming at expanding drug indications and discovering potential adverse events. The real-world data is fraught with missing values. Literature addressing different strategies for dealing with missing data in such a situation is scarce. Using machine learning (ML) algorithms for predicting therapeutic outcomes of PD-1/PD-L1 Inhibitors has attracted attention. However, training a predictive model usually requires imaging or biomarker information, which is rarely available in the post-marketing surveillance data. To address these challenges, we propose an ML-aided framework to predict the outcomes of Anti-PD-1 therapy for gynecological malignancy on a dataset with 117 patient samples, treated by Camrelizumab (with 50 patient samples), Sintilimab (44), and Toripalimab (23). Four therapeutic outcomes, including Response Evaluation Criteria in Solid Tumours (RECIST), organ adverse effect (AE), general AE, and death, are predicted. The proposed framework feeds the dataset into a learning pipeline consisting of imputation, feature engineering, model training, ensemble learning, and model selection to generate the final predictive model. We conduct experiments to justify several critical design choices, such as the specific feature engineering strategies and the SMOTE over-sampling technique. The final model for each learning task is selected from a large pool of model candidates based on a joint consideration of accuracy and F1. Moreover, we conduct thorough and visualized model analysis and gain a deeper understanding of model behavior and feature importance. The results, analysis, and findings demonstrate the superiority of the proposed learning-aided framework.

INDEX TERMS Post-marketing surveillance, machine learning, gynecological cancer.

I. INTRODUCTION

Conducting a post-marketing surveillance study is significant for evaluating the efficacy and safety of a drug in clinical practice [1]. Recent years have witnessed the development and application of PD-1/PD-L1 inhibitors in cancer therapy [2], [3]. The PD-1/PD-L1 inhibitors work by blocking the interaction of PD-L1 on a tumor cell with PD-1 on a T-cell, allowing the immune system to start functioning and attack tumor cells [4]. Understanding the human body response to the PD-1/PD-L1 therapy via post-marketing surveillance is critical to improving treatment programs and boosting

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

research and development of PD-1/PD-L1 inhibitors in the long run.

Worldwide, around 1.5 million women are diagnosed yearly with a gynecological malignancy [5]. In the UK and USA, ovarian cancer is the fifth most common cause of cancer death amongst women; endometrial cancer is the fourth most common cancer for women and the most common gynecological cancer in developed countries; another common cancer type, cervical cancer, has become mostly preventable in the developed world due to effective screening and vaccination [6]. The efficacy of PD-1/PD-L1 inhibitors has been demonstrated in the treatment of gynecological cancer, with a high response rate and a tolerable side effect profile [7]. In a study by Meng *et al.* [8], PD-L1 expression was found in 60.82% (59/97) of the patients with cervical cancer. In another study [9], 123 patients with high-grade serous ovarian carcinoma were treated with PD-L1 inhibitor and chemotherapy, and 87 (71%) and 29 (24%) were evaluated as a complete and partial response, respectively. Meanwhile, side effects caused by upregulation of the immune system have shown on several organ systems, including the skin, the gastrointestinal tract, the liver, and the endocrine system [6] for patients receiving anti-PD-1 therapy. Fatigue was reported by 16–37% of patients [10]. 30–40% of patients presented dermatologic toxicities [11], most commonly a maculopapular rash. Diarrhea and colitis have also been reported but not common, occurring in 1–3% of patients [12]. Less than 5% of patients reported immune-related hepatitis with an asymptomatic elevation of liver transaminases [10]. Compared to CTLA-4 inhibitors, Hypophysitis is less common with PD-1/PD-L1 inhibitors, only observed in 1–6% of patients; thyroiditis, on the other hand, is more common, seen in 6–20% of patients [10].

Numerous prior studies have employed conventional statistical methods to analyze the efficacy and safety of anti-PD-1 therapy of gynecological malignancy [7], [13]–[16]. These methods allow clinical practitioners to uncover the basic statistical facts in the data gathered through post-marketing surveillance. Recent advances in artificial intelligence and machine learning (ML) have seen great success in various industries. In the sub-field of PD-1 inhibitor application, recent studies have employed ML techniques to mine informative patterns hidden in the surveillance records and build models to predict the anti-PD-1 therapy outcomes with statistically validated confidence. Existing cases include non-small cell lung cancer (NSCLC) [17]–[19], metastatic melanoma [20], and intrahepatic cholangiocarcinoma (ICC) [21] . However, ML-based analytical methods have not been seen in the area of anti-PD-1 therapy for gynecological malignancy.

Training an accurate and robust predictive model requires a high-quality dataset [27]–[29]. By high quality, we mean a large data volume, accurate labels, and high integrity. However, the dataset used in this study is small, with only 117 samples, and there are missing values. The only high-quality criterion met is the data labels, which are the therapy outcomes gathered from patients in the real world. The small and incomplete dataset brings a unique challenge for predictive learning. Thus, a crucial question we need to answer is that whether our models can learn meaningful patterns to produce accurate results in this challenging scenario. Similar challenges have been faced by prior studies. Various imputation methods have been developed to deal with the missing values [30], [31]. The cases of the small dataset have been seen in the studies of brain disorders [31] and Alzheimer's Disease [28]. When obtaining more data is not viable, a common strategy is data augmentation [32], [33], which can be used to create synthetic samples to increase the training data size and rebalance class distribution [34].

From the perspective of learning strategy, numerous methods have been developed from both the data and model aspects. It is crucial for a learning algorithm to mine informative patterns from a given set of features. Common strategies for feature handling include normalization [35], feature enhancement [36], and feature selection [37]. From the model side, ensemble learning has shown superior predictive performance in a wide spectrum of ML systems [38], [39] due to its ability to strategically aggregate a collection of single learning models and achieve performance gains. However, our investigation shows that these common learning strategies are rarely seen in the literature that adopts ML to predict outcomes of anti-PD-1 treatment.

Table 1 displays a list of nine relevant studies of ML-based outcome prediction for anti-PD-1 therapy, sorted by the year of publication. We compare these efforts in seven dimensions, including cancer type, dataset size, the number of predicted outcomes, the usage of imputation, feature enhancement, data augmentation, feature selection, and ensemble learning. It is noted that NSCLC and melanoma are the common cancer types for this kind of study, appearing in seven papers. Also, the datasets adopted by most studies (seven out of nine) are small, with a size from 56 to 250. There is a large dataset with over 95k samples, an aggregated one downloaded from the US Food and Drug Administration Adverse Event Reporting System [24]. All nine studies only predict one or two outcomes of the anti-PD-1 therapy. None consider imputation, although missing values did present [22], [24], [26]. In [26], lack of handling missing values was listed as a study limitation. In [24], samples with missing values were not used. Paper [22] claims that missing data were minimal due to robust documentation and follow-up information. It is also observed that feature enhancement, a technique to create new features based on existing ones, is rarely used [23], while feature selection is more commonly considered [17], [18], [21], [22], [24]–[26]. In addition, despite working on a small dataset, none of the reviewed studied considered data augmentation, and none employed ensemble learning. Our investigation indicates that, for the task of outcome prediction of anti-PD-1 therapy, methods to handle missing values and more learning strategies remain to be explored. Our study aims to fill this methodological gap.

In this paper, we propose an ML-based framework to perform predictive analysis on the outcomes of anti-PD-1

TABLE 1. A comparative review of literature. Acronyms in table include predicted outcomes (P.O.), imputation (Imp.), feature enhancement (F.E.), data augmentation (D.A.), feature selection (F.S.), ensemble learning (E.L.), non-small cell lung cancer (NSCLC), and intrahepatic cholangiocarcinoma (ICC).

Work	Cancer Type	D.S.	# P.O.	Imp.	F.E.	D.A.	F.S.	E.L.
Naik (2018) [22]	Melanoma brain metastases	56	Two	✗	✗	✗	✓	✗
Song et al. (2019) [17]	NSCLC	250	One	✗	✗	✗	✓	✗
Yan et al. (2019) [23]	NSCLC and melanoma	1,055	One	✗	✓	✗	✗	✗
Lewinson et al. (2020) [20]	NSCLC and melanoma	142	One	✗	✗	✗	✗	✗
Zhang et al. (2020) [21]	ICC	98	One	✗	✗	✗	✓	✗
Wiesweg et al. (2020) [18]	NSCLC	122	One	✗	✗	✗	✓	✗
Yang et al. (2020) [24]	Multiple types	95,918	One	✗	✗	✗	✓	✗
Johannet et al. (2021) [25]	Advanced melanoma	151	Two	✗	✗	✗	✓	✗
Ahn et al. (2021) [26]	NSCLC	192	One	✗	✗	✗	✓	✗
Ours	Gynecological	117	Four	✓	✓	✓	✓	✓

therapy for patients with gynecological cancers, using a post-marketing surveillance dataset with 117 patient profiles collected between Sep 2019 and Feb 2021. We focus on four therapeutic outcomes as the prediction targets, including Response Evaluation Criteria in Solid Tumours (RECIST), organ adverse effect (AE), general AE (namely side effect), and death. The learning tasks considered in this study consist of both multi-class and binary classification problems, which can be processed through a custom learning pipeline. To handle missing values in the dataset, we employ six candidate imputation methods to fill the blanks. After imputation, the dataset undergoes a series of feature engineering operations, including encoding, normalization, and interaction. The last operation, namely feature interaction, is a feature enhancement method that creates new features based on existing ones. The enhanced dataset is then divided into training and test sets. For training, a total of fifteen candidate models are considered. For each learning task, each candidate model is trained with the Synthetic Minority Over-sampling Technique (SMOTE) [40] for data augmentation, and the recursive feature elimination (RFE) [41] for feature selection. Each trained model is then tuned to find the optimal set of hyperparameters within the pre-defined search space. The top N tuned models are sent to an ensemble learning module to produce ensemble models. Finally, both ensemble and base models are evaluated on the test set to identify the best model for the task.

We conducted extensive experiments to validate the effectiveness of the proposed learning pipeline. Results can justify several crucial design choices, such as the feature engineering and selection strategies and the data augmentation method. Furthermore, through a thorough model analysis, we gain a deeper understanding of the learning process and how the features can jointly affect the performance of our models.

The rest of this paper is organized as follows. Section II describes the dataset details. Section III discusses the technical details of the proposed analytical framework. Section IV presents the evaluation results and model analysis. We provide an extensive discussion of this work in Section V.

II. DATASET

A. BASIC INFORMATION

The study investigated a gynecological cancer post-marketing surveillance dataset with 117 patient samples collected from Sep 2019 to Feb 2021. The study also compares three PD-1 inhibitors, including Camrelizumab [42] (with 50 patient samples), Sintilimab [43] (44 samples), and Toripalimab [44] (23 samples). Table 2 shows the columns of the dataset. The columns are referred to as features in ML tasks, and we use the term feature in the rest of this paper. A total of 15 features were collected from the patients, including age, history of hepatitis, tumor type, PD-1/PD-L1 inhibitor, lines of therapy, chemotherapy, radiotherapy, targeted therapy, start time, medication cycle, white blood cells (WBCs), aspartate aminotransferase (AST), alanine aminotransferase (ALT), thyroid-stimulating hormone (TSH), and tubercle bacillus (TB). Based on the data type, the features can be divided into binary, nominal, ordinal, time, and continuous data. Also, the dataset is incomplete due to missing data. Specifically, three features, including age, lines of therapy, start time, and medication cycle, suffer mild missing, i.e., the missing rate is less than 10%. Six features, including the history of hepatitis, WBCs, AST, ALT, TSH, and TB, suffer severe missing, i.e., the missing rate is more than 50%.

B. FEATURE STATS

We uncover the statistical information of individual features as follows:

- The age data can be divided into six bins from the 20-year-old age group through the 70-year-old age group, as shown in Figure 1 (a). The age data points follow a slightly left-skewed distribution, with 49 samples (the most) in the 50-year-old group and five samples (the least) in the 20-year-old group. The mean and median of all data points are 55.49 and 56, respectively.
- Table 3 presents the stats information of the gynecological tumor type in the dataset, where the three most common types are cervical cancer (63 cases), ovarian cancer (20 cases), and endometrial cancer (20 cases). The total count of the rest of the cancer cases is only 14.
- The gathered values for the lines of therapy are either one or two, with three missing data points, as shown in Figure 1 (b). Among the samples, most patients (74%) were in their second line of therapy, meaning that an earlier treatment has been conducted, and tumor recurrence has occurred.

TABLE 2. PD-1/PD-L1 Inhibitor application in gynecological cancer dataset.

Feature	Type	% Missing	Mean	STD	Min	Max
Age	Ordinal	0.85%	55.49	10.52	26	77
History of hepatitis	Binary	52.54%	-	-	0	1
Tumor type	Nominal	0.00%	-	-	-	-
PD-1/PD-L1 Inhibitor	Nominal	0.00%	-	-	-	-
Lines of therapy	Ordinal	2.54%	1.73	0.44	1	2
Chemotherapy	Binary	0.00%	-	-	0	1
Radiotherapy	Binary	0.00%	-	-	0	1
Targeted therapy	Binary	0.00%	-	-	0	1
Start time	Time	5.08%	-	-	9/4/2019	12/31/2020
Medication cycle	Ordinal	5.93%	4.62	4.38	1.00	20.00
WBCs	Continuous	54.24%	5.90	3.55	1.60	17.40
AST	Continuous	54.24%	20.09	11.06	5.00	74.00
ALT	Continuous	54.24%	19.26	12.16	1.00	69.00
TSH	Continuous	85.59%	1.91	1.63	0.01	6.21
TB	Continuous	54.24%	8.93	5.75	1.50	41.80

TABLE 3. Gynecological cancer type in the dataset.

	Sintilimab	Toripalimab	Camrelizumab	Grand Total
Ovarian cancer	11	2	7	20
Round Ligament cancer	1	0	0	1
Vulvar cancer	0	0	1	1
Endometrial cancer	12	2	6	20
Uterine Sarcoma	0	0	2	2
Cervical cancer	18	15	30	63
Trophoblastoma	0	1	2	3
Choriocarcinoma	2	2	0	4
Fallopian tube cancer	0	1	0	1
Vaginal cancer	0	0	2	2

- Regarding the therapy program, 40 patients received single anti-PD-1 therapy, and the rest 77 patients received a joint therapy program with PD-1 and chemotherapy/radiotherapy/targeted therapy. Figure 1 (c) shows the stats of the therapy type grouped by PD-1. It is noted that the three therapy types are not mutually exclusive. In other words, patients could receive one or more types of therapy. For example, one patient in the dataset received all three therapy combined with Sintilimab.
- Figure 1 (d) shows the histogram of medication cycle grouped by PD-1. It can be observed that most patients (71%) have taken the medication for one through four cycles. Also, we observe that the twelve patients with over ten medication cycles were all given Camrelizuma.

C. OUTCOME

The therapy outcome is evaluated in four dimensions, including the RECIST, death, organ AE, and general AE. A breakdown of the four aspects is as follows:

- RECIST is a guideline used to evaluate the response of therapy on lesions. There are four levels of response, including complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). Table 4 displays the stats of the RECIST grouped by PD-1. The combined percentages of CR and PR for the three PD-1 inhibitors are 52% (Camrelizuma), 65.91% (Sintilimab), and 82.61% (Toripalimab), respectively, indicating an increment of efficacy.
- A total of 15 death cases are counted in the dataset, as shown in Table 5. The death rates for Camrelizuma, Sintilimab, and Toripalimab are 18%, 13.64%, and 0%, respectively.
- Table 5 also shows that the percentages of organ AE occurrences for Camrelizuma, Sintilimab, and Toripalimab are 64%, 50%, and 43.48%, respectively, and the percentages of general AE for Camrelizuma, Sintilimab, and Toripalimab are 18%, 47.73%, and 17.39%.

D. THE LEARNING TASKS

There are four learning tasks, which are to predict the four types of outcomes, including RECIST, death, organ, and general AE, using the data points in the dataset. Specifically, the latter three belong to binary classification, as the targets are binary. The first task, RECIST prediction, contains four classes, making it a multi-class classification problem. Since dividing the dataset into four classes would further reduce the sample size for each class, it is harder to train an accurate model [37]. Thus, in addition to the original multi-class classification task, we add an auxiliary task by merging CR and PR to make a C&P R class and merging SD and PD to create an S&P D class. The four-class classification problem is then reduced to a binary classification problem (referred to as binary RECIST). A major challenge we face is the imbalanced sample distribution across the dataset. Figure 3 shows the class distributions for the five learning tasks.

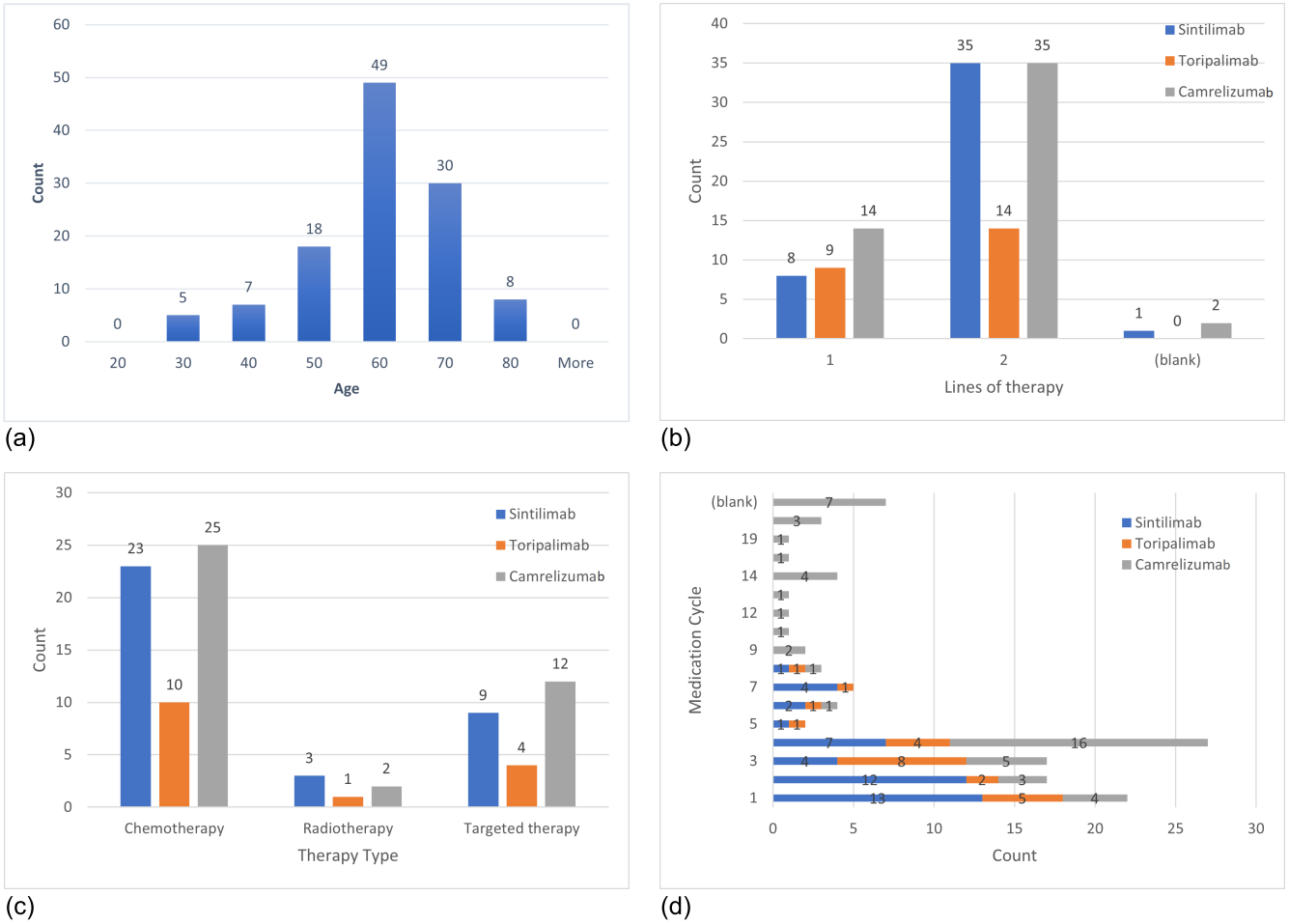

FIGURE 1. Stats of individual features: (a) age, (b) lines of therapy, (c) therapy type, and (d) medication cycle.

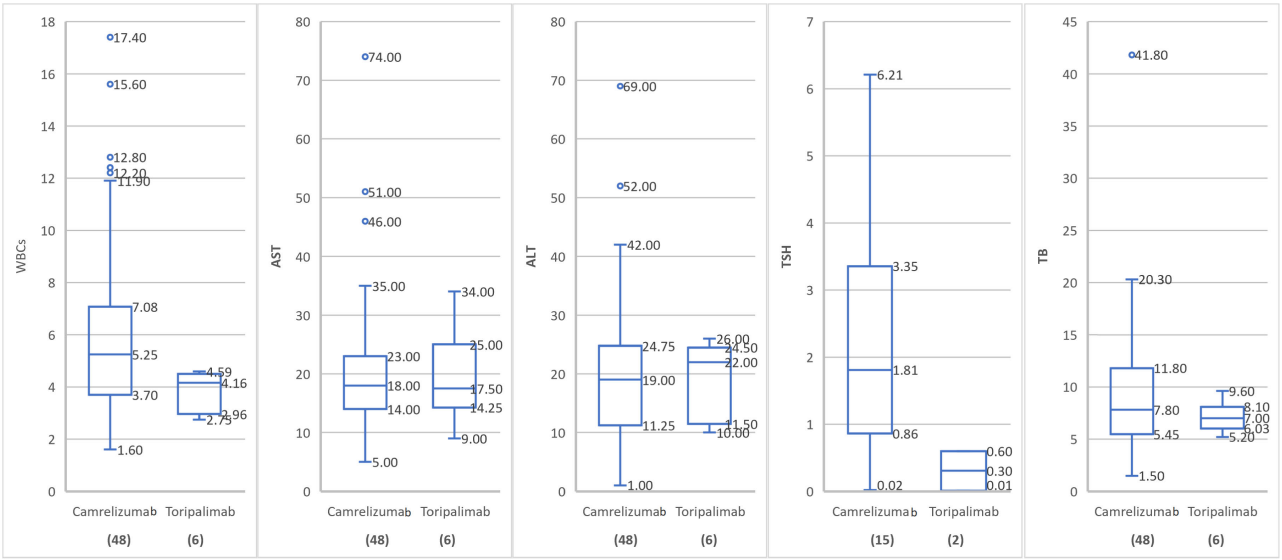

FIGURE 2. Laboratory results of WBCs, AST, ALT, TSH, and TB (left to right). The number in braces under the PD-1 inhibitor is the sample size.

For the binary learning tasks, the percentages of the major classes of binary RECIST, general AE, organ AE, and death are 62.71% (C&P R), 70.34% (general AE not present), 54.24% (organ AE present), and 87.29% (alive), respectively.

TABLE 4. RECIST stats.

PD-1/PD-L1	Size	CR	%	PR	%	SD	%	PD	%
Camrelizumab	50	15	30.00%	11	22.00%	7	14.00%	12	24.00%
Sintilimab	44	8	18.18%	21	47.73%	9	20.45%	6	13.64%
Toripalimab	23	11	47.83%	8	34.78%	1	4.35%	3	13.04%

TABLE 5. Death and AE stats.

PD-1/PD-L1	Size	Death	%	Organ AE	%	General AE	%
Camrelizumab	50	9	18.00%	32	64.00%	9	18.00%
Sintilimab	44	6	13.64%	22	50.00%	21	47.73%
Toripalimab	23	0	0.00%	10	43.48%	4	17.39%

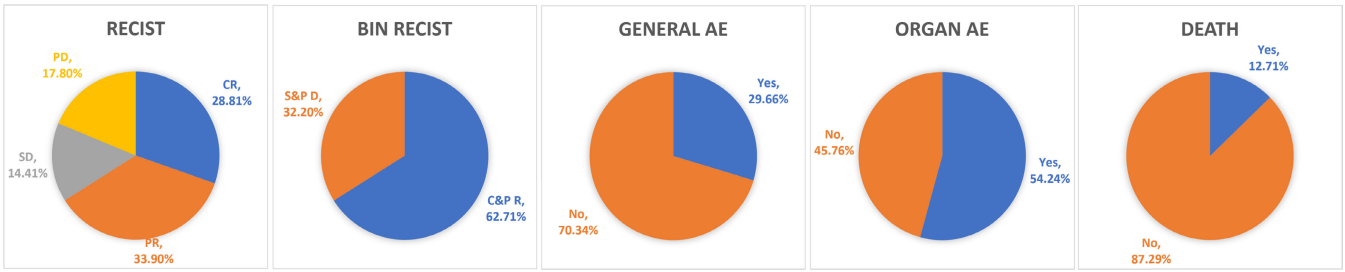

FIGURE 3. Sample distribution for each learning task.

For the RECIST task, samples in the four classes are also not equally distributed. Since our dataset is tiny, the imbalanced data distribution leads to insufficient training samples for minority classes, affecting the learning performance.

Before training, it is necessary to obtain a general sense of the potential impact of each individual feature on the learning outcomes. We provide a correlation heatmap in Figure 4 for all columns of the dataset, including features and outcomes. Specifically, we employed Pearson's r [28] as a measure of linear correlation between two columns of data to generate a matrix, where each element is between -1 and 1. In other words, the higher the Pearson's r, the stronger positive correlation presented by two sets of data, and vice versa. It is noted that before the calculation of the correlation matrix, we applied a scheme called one-hot encoding to binarize the two nominal categorical features, namely PD-1 and tumor type. This way, each value of these two features is transformed into an additional binary feature based on its occurrence in the original feature. For example, the original feature PD-1 becomes three binary features named PD-1_Camrelizuma, PD-1_Sintilimab, and PD-1_Toripalimab. The reason for one-hot encoding is that it allows the correlation calculation and the following learning algorithms to better understand the relationship between each individual value of the categorical feature and the outcome. The key observations of Figure 4 are as follows:

- The death outcome has a relatively high positive correlation with Tumor type_Ovarian cancer (with a Pearson's r 0.18), Age (0.15), and PD-1_Camrelizumab (0.13). The top three negatively correlated features are PD-1_Sintilimab (-0.19), radiotherapy (-0.18), and therapy lines (-0.17). Also, death is highly correlated with another outcome, RECIST (-0.4).
- The RECIST and bin RECIST are highly correlated since bin RECIST is a looser setting of RECIST. In addition, the top three positively correlated features are Tumor type_Cervical cancer (0.22), Tumor type_Choriocarcinoma (0.22), and Tumor type_Trophoblastoma (0.19), and the top three negatively correlated features are Tumor type_Ovarian cancer (-0.47), Age (-0.35), and Therapy lines (-0.27). It is noted that the four classes CR, PR, SD, and PD are encoded as 4, 3, 2, 1, respectively. Therefore, a high Pearson's r indicates that the feature has a positive effect on the treatment.
- The top three positively correlated features with organ AE are Medication cycle (0.27), Chemotherapy (0.21), and PD-1_Camrelizumab (0.21), and the top three negatively correlated features are Starting time (-0.17), TSH (-0.15), and Tumor type_Cervical cancer (-0.14).
- The top three positively correlated features with general AE are PD-1_Sintilimab (0.3), Tumor type_Endometrial cancer (0.17), and History of hepatitis (0.16), and the top three negatively correlated features are PD-1_Camrelizumab (-0.22), Medication cycle (-0.16), and TSH (-0.12).

III. A MACHINE LEARNING-BASED ANALYTICAL FRAMEWORK

A. OVERVIEW OF FRAMEWORK

Figure 5 shows an overall learning framework of this study. First, the original dataset is imputed to fill the missing values.

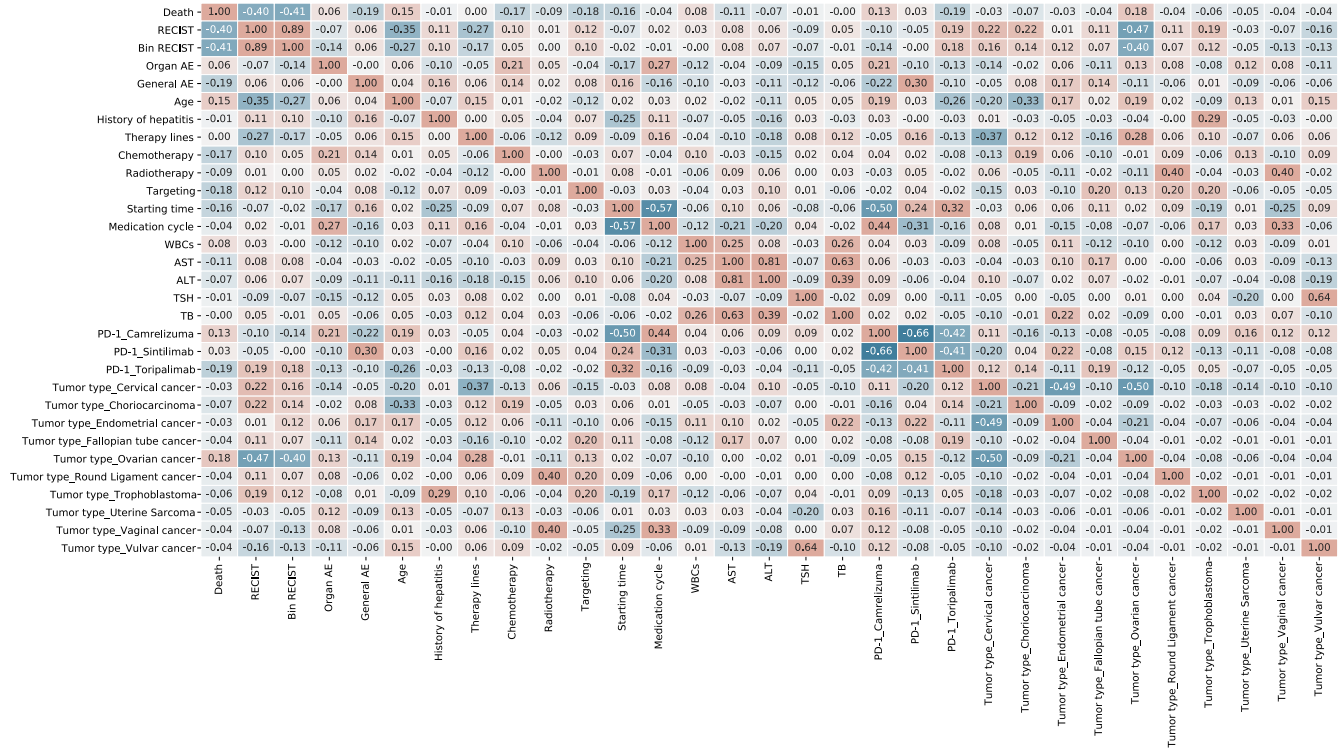

FIGURE 4. Correlation heatmap.

We evaluate six imputation methods with different designs. The completed dataset then undergoes a feature engineering module that consists of encoding, normalization, and feature interaction. After feature engineering, we obtain an expanded feature set from the 15 raw features. The dataset is then divided into a training and test set. For each learning task, we train each candidate model on the training set, using SMOTE to generate synthetic samples and RFE for feature selection. The trained models (referred to as base models) then undergo a process of hyperparameter tuning. The top models are sent to an ensemble learning module to create ensemble models. Finally, the tuned base models and the ensemble models are evaluated on the test set to determine the best model for the learning task.

B. DATA IMPUTATION

Since the raw dataset is incomplete, we employ data imputation to fill the missing values. The following seven imputation methods are adopted in this study.

- A simple imputer applies a pre-defined strategy to fill the missing values. Common strategies include filling with mean, median, constant, and most frequent. In this study, we chose the median-filling strategy since it outperformed other strategies in the learning tasks. The simple imputer has been widely used as a baseline in prior studies [45], [46].
- An iterative imputer fills the missing values by treating each feature as a target and using all other features to train a regression model. The process operates round-robin fashion until all features with missing values are imputed. In this study, we chose the Bayesian ridge model [47] for regression. It is noted that the iterative imputer relies on the simple imputer to initialize the dataset, and the missing values are filled with placeholders that go through an iterative procedure for refinement.
- The K-Nearest Neighbors (KNN) [48] imputer first identifies the K-nearest neighbors for each data point of the dataset; then, a missing value is imputed with the mean value belonging to the K-nearest neighbors.
- The Multiple Imputation by Chained Equations (MICE) [49] is similar to the iterative imputation but implemented differently. MICE creates multiple copies of the dataset and applies the MICE procedure to each dataset copy. Lastly, the filled values in all copies are pooled to form the final completed dataset.
- The MissForest imputation [50] fits a random forest for each feature with missing values using the observed part to predict the missing part. This process operates iteratively until a stop condition is satisfied.
- The Generative Adversarial Imputation Nets (GAIN) [51] trains two networks, including a generator to impute the missing values based on the observed ones and a discriminator to distinguish observed and imputed data. The two networks are trained and optimized alternatively to improve the imputation quality

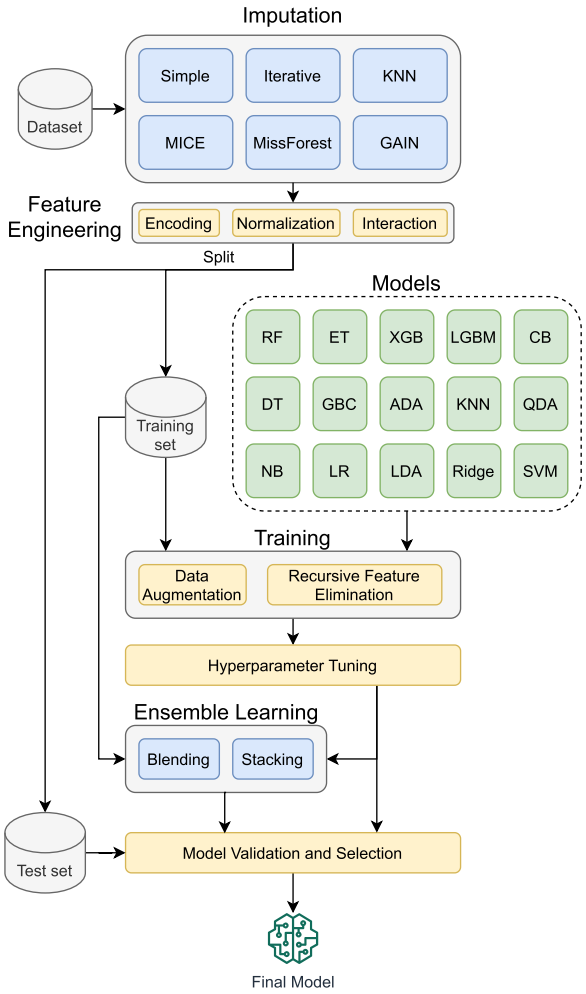

FIGURE 5. The overall learning framework. We experiment with six imputation methods, including simple, iterative, K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), MissForest, and Generative Adversarial Imputation Nets (GAIN). Also, the set of candidate models adopted in this study includes Random Forest (RF), Extra Trees (ET), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), CatBoost (CB), Decision Tree (DT), Gradient Boosting Classifier (GBC), Ada Boost (ADA), K-Nearest Neighbors (KNN), Quadratic Discriminant Analysis (QDA), Naive Bayes(NB), Logistic Regression(LR), Linear Discriminant Analysis (LDA), Ridge, and Support vector machine (SVM).

until convergence. When the training is done, the generator has learned the distribution of the data points and can generate realistic samples that are difficult to be detected by the discrminator as generated ones.

C. FEATURE ENGINEERING

We adopted three techniques to engineer the raw features. First, as mentioned in Section II-D, we applied one-hot encoding [52] which can binarize the categorical features so that each feature value can be treated individually during training and make a dedicated impact on the prediction target. Also, binarized categorical features are easier to interpret in model analysis. Second, we normalized the continuous features using the z-score [35], which is computed as $z = \frac{x-\mu}{\sigma}$, where x is the raw feature value, and μ and σ refer to the mean and standard deviation of the population, respectively. Lastly, we employed feature interaction [53], which creates new features with pairwise multiplication between the raw features. A broader design choice would be to generate polynomial features [54] that consist of all polynomial combinations of the raw features. However, we found that the pairwise multiplication consistently outperformed the polynomial feature generation by 2-6% in our experiments. Thus, we only kept pairwise multiplication for feature interaction, which reduces the model complexity, shortens training time, and boosts the model performance.

D. LEARNING FROM DATA

Once the dataset is complete, we can start the process of model training. This subsection describes the major steps of the learning pipeline.

1) MODEL TRAINING

We chose fifteen base models that commonly appear in both literature and industrial applications. A brief description of each model is provided as follows.

- The **Decision Tree** Classifier (DT) [55] is a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
- The **Random Forest** Classifier (RF) [56] is an ensemble learning method that operates by combining a multitude of DTs during training and outputting a class via voting. Combining a group of weak learners to form stronger learners is the main improvement in Random Forest compared with conventional DT structure after introducing an ensemble approach [57]. A divide and conquer approach is utilized in ensemble methods to achieve an improvement in algorithm performance. The average of these uncorrelated trees can be obtained in order to achieve a variance reduction [58] which can make the final results less variable and more reliable [59].
- The **Extra Trees** Classifier (ET) [60] fits a number of randomized DTs (a.k.a. extra-trees) on various sub-samples of the dataset.
- The **Light Gradient Boosting Machine** (LightGBM) [61] grows trees leaf-wise (best-first), which is more efficient than other tree-based boosting methods.
- The **Extreme Gradient Boosting** (XGB) introduced as a tree-based method, is proposed by Chen in 2014 [62]. As a scalable and accurate implementation of gradient boosted trees, optimized computational speed and model performance can be achieved. Reduction in the overfitting effect can be realized by XGB utilizing a regularization term compared to gradient boost, which is contributed to better prediction and much faster computational run times [63].
- The **CatBoost** Classifier [64] is a ML algorithm that uses gradient boosting on DTs.
- The **Gradient Boosting** Classifier (GBC) [65] is an ML framework that treats boosting as an optimization

problem that aims to minimize a loss function of the model by adding weak learners via a procedure similar to gradient descent. DTs are usually adopted as the weak learners, which are built in a greedy manner by splitting nodes with the highest purity score. The algorithm is a stage-wise additive process, meaning that a weak DT leaner is added one at a time, and the existing DTs in the model remain unchanged. The prediction of the new tree added to the predictions of the existing models attempting to improve the model performance. The training stops when a pre-defined number of trees are added, or the loss reaches an acceptable level.

- The **Adaptive Boosting** Classifier (ADA) [66] takes the output of other learning algorithms ('weak learners') to obtain a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.
- The **K Nearest Neighbors** Classifier (KNN) [67] uses a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.
- The **Quadratic Discriminant Analysis** (QDA) [68] is a statistical classifier that uses a quadratic decision surface to separate measurements of two or more classes of objects or events.
- The **Naive Bayes** (NB) [69] model is a set of supervised learning algorithms based on the Bayes' theorem with the assumption of conditional independence between every pair of features given the value of the class variable.
- The **Logistic Regression** (LR) Originally proposed by Cox in 1958 [70], linear discriminants is commonly involved in Logistics Regression as a traditional classification algorithm. The given input can obtain a probability as a primary output belonging to a certain class. Then a linear boundary can be established to separate the input space into two regions by the model based on the value of probability. Linearly separable classes can be solved easily by logistics regression as one of the most extensively used classifiers. This logistic function is a common "S" shape (sigmoid curve).
- The **Linear Discriminant Analysis** (LDA) [68] is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using the Bayes' rule. The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix. The fitted model can also be used to reduce the dimensionality of the input by projecting it to the most discriminative directions.
- The **Ridge** Classifier [71] first converts binary targets to -1, 1 and then treats the problem as a regression task, optimizing the same objective as above. The predicted class corresponds to the sign of the regressor's prediction.
- The Linear Kernel **Support vector machine** (SVM), a commonly used discriminative classifier, is designed to assign new data samples to one of two possible categories and was initially proposed in 1995 by Vapnik and Cortes [72]. A hyperplane separating the n-dimension data into two classed is the basic idea of SVM, and the maximized geometric distance to the nearest data points can be obtained by the hyperplane, that is, support vector. Notably, similar results can often be yielded by practical linear SVM and logistics regression [73]. Apart from linear classification, a kernel method that can perform efficiently in non-linear classification is introduced by SVM. The attributes can be transferred into a new feature space where the data is separable by kernel, a feature mapping methodology.

2) DATA AUGMENTATION

To alleviate the negative effect of the imbalanced dataset, we adopted the SMOTE [40] to synthesize new data points from the minority class. Essentially, SMOTE is a data augmentation method aiming to re-balance the dataset by creating equally distributed classes. At a high level, SMOTE works by choosing a pair of samples that are close in the feature space and drawing a line on the two samples, and then a new data point can be drawn along the line. This way, more samples in the minority class can be synthesized to balance the dataset. Compared to other oversampling methods, such as Edited Nearest Neighbours, SMOTE is efficient and easy to implement [74], [75].

3) RECURSIVE FEATURE ELIMINATION

The RFE [41] aims to find a subset of features among all available features to train the final model. It works by training a full model with all features and ranking these features by importance. The training process repeats with the insignificant features gradually removed. Cross-validation (CV) is adopted to evaluate a model's performance given a set of selected features, and the optimal feature set is kept when RFE finishes. Compared with other feature selection methods such as InfoGain [76] and ReliefF [77], RFE is suitable for dataset with small size and high input noise [78]. For our case, we have a small dataset with input noise introduced by missing values, making RFE a decent choice.

4) HYPERPARAMETER TUNING

We conduct a randomized search on hyperparameters. The random parameter optimization scheme works by randomly searching over the pre-defined parameter space, where each parameter setting is sampled from a distribution over possible parameter values [79]. Compared to the exhaustive grid search, randomized search has two strengths [80]. First, a budget can be reserved to control the computational resource and load. Second, adding new parameters has little influence on the training efficiency.

5) ENSEMBLE LEARNING

Ensemble learning is a technique that strategically aggregates a set of predictive models to boost predictive performance [38], [39]. In this study, we consider two ensemble methods, blending and stacking. The idea of blending [81] is to combine different ML algorithms and use a majority vote or the average predicted probabilities in case of classification to predict the outcome. On the other hand, the stacking method [81], [82] treats the outputs of the base estimators as input to train a second-level model on top of a selected meta estimator. Ensemble learning usually demonstrates superior performance than single models because the strategy of model aggregation has the potential of finding more distinguishable patterns, which may not be seen by a single model. However, the performance of an ensemble could be affected if the final prediction is controlled by some weak models that are the majority and too weak to give correct predictions, leading to an accuracy degradation even if a strong model is correct all the time.

E. MODEL EVALUATION

Four metrics are evaluated for each learning task in this study, including accuracy, precision, recall, and the F1 score. Given the imbalanced dataset, F1 is a better metric than accuracy in binary classification where positive samples make the minority class [83]. Precision and recall are also crucial. The former reflects the number of false alarms; the higher the precision, the fewer false alarms. The latter tells the number of missed positive samples; the higher the recall, the fewer positive samples missed. A large precision-recall gap should be avoided since it indicates that a model focuses on a single metric, while a model should really focus on optimizing F1, the harmonic mean of precision and recall. Let TP, FN, TN, and FP denote the number of true positives, false negatives, true negatives, and false positives, respectively, we can then calculate accuracy, precision, recall, and F1 as follows.

$$accuracy = \frac{TP + FN}{TP + FN + TN + FP} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETTING

All experiments of this study were implemented using Python 3.7.0. The learning algorithms were provided by Scikit-learn 0.24 [84]. The charts were plotted using Microsoft 365 Excel, Matplotlib 3.4.2, and Seaborn 0.11. The 112 samples were divided into a training (78) and test (34) set in the ratio of 7:3. A five-fold CV was conducted for RFE and hyperparameter tuning to determine the optimal set of features and hyperparameters, respectively.

B. RECURSIVE FEATURE ELIMINATION

We adopt RFE to perform feature selection for each model. As shown in Figure 6, the horizontal axis is the number of features selected, and the vertical axis is the CV F1. For this specific model, the optimal feature set contains 28 features, achieving an F1 of 0.694. In our study, RFE is used to determine the set of features used to train a model.

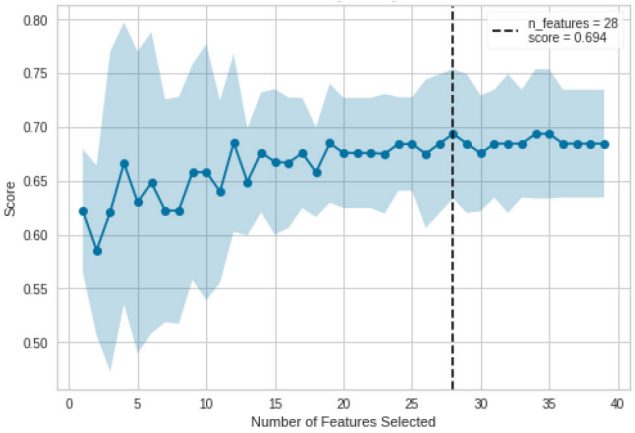

FIGURE 6. An example of using RFE for feature selection.

C. HYPERPARAMETER TUNING

We adopt the randomized hyperparameter search algorithm implemented by Scikit-learn [84]. For each model, a pre-defined set of parameters is provided by Scikit-learn, and the search algorithm randomly and repeatedly sample a parameter setting for evaluation until a stop condition is met. For each model, we set a maximum iteration number of 500 and choose the parameter setting that yields the best CV F1.

D. RESULTS

The main results are reported in Tables 6 through 10. For each learning task, we evaluated fifteen models for each imputation method and only showed the performance of the top-performing model per imputation method, making six rows of results, displayed in the top section of each table. The bottom section of each table, on the other hand, shows the performance of the two ensemble models built with the top five models imputed by the best imputer. For example, in Table 6, the best imputer was the simple imputer, since the top model imputed by the simple imputer presented the highest F1 than the top models imputed by other imputers. Therefore, we took the top five models imputed by the simple imputer (the other four not showing in the table) to build the ensemble models, as shown in the last two rows of Table 6. For the convenience of presentation, we use a format of imputation-model to denote a model imputed by an imputation method, e.g., simple-XGB. In addition to the main results, we also report 1) an ablation study that presents the results of each prediction task under different training settings, 2) the confusion matrix, and 3) the feature importance chart. The latter two

are obtained from the best models. We present a breakdown of these results in the following subsections.

1) PREDICTION OF DEATH

a: ABLATION STUDY

Figure 7 presents the box plots of the five-fold CV F1 scores of all fifteen models for the death prediction task under different training settings. Specifically, we started with a base setting with only one-hot encoding to transform categorical features and incrementally applied normalization, feature interaction, SMOTE, RFE, and hyperparameter tuning to the base setting, resulting in six settings. We observe a clear and consistent rising trend of the F1 scores from left to right. In theory, normalization allows the learning algorithm to converge quickly; feature interaction uncovers a set of new features that may be more indicative of the prediction target; SMOTE can sufficiently address the imbalanced class issue by feeding generated samples to the learning algorithm; RFE can identify the best set of features; hyperparameter tuning can help determine a better set of hyperparameters. These strategies jointly lead to a remarkable performance gain, elevating the best CV F1 from 0.41 (base setting) to 0.89 (with all strategies, referred to as the full setting). The results of this experiment validate these hypothetical pros.

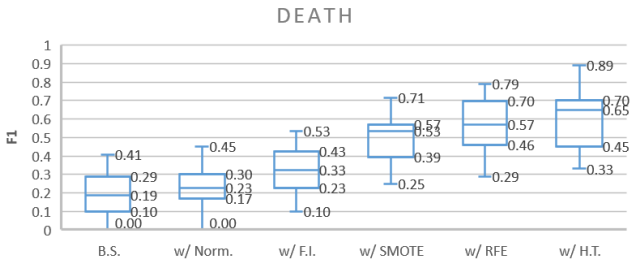

FIGURE 7. Results of death prediction under different training settings. Acronyms in the chart include base setting (B.S.), normalization (Norm.), feature interaction (F.I.), and hyperparameter tuning (H.T.)

b: OVERALL PERFORMANCE

The results of death prediction on the test set are reported in Table 6. This task is challenging due to the imbalanced class issue. A model can easily obtain an accuracy of 87% by predicting every sample as negative; however, this would lead to an F1 of 0% since no TP is counted. It can be seen that the simple-XGB model was the best one, with the highest accuracy (0.9706), recall (1), precision (0.75), and F1 (0.8571).

c: CONFUSION MATRIX

Figure 8 shows the confusion matrix of the best model, namely Simple-XGB, for death prediction on the test set, which consists of 31 negative cases and three positive cases. The Simple-XGB model was able to correctly classify 33 out of 34 samples, with only one negative case misclassified.

TABLE 6. Result of death prediction on the test set. The highest score of each metric is marked in bold-face.

Imputation	Model	Accuracy	Recall	Precision	F1
Simple	XGB	**0.9706**	**1**	**0.75**	**0.8571**
MICE	CB	0.9118	0.3333	0.5	0.4
Iterative	ADB	0.9118	1	0.5	0.6667
KNN	SVM	0.8529	0.3333	0.25	0.2857
MissForest	RF	0.9118	0.3333	0.5	0.4
GAIN	CB	0.9118	0.6667	0.5	0.5714
Simple	Blender	0.9118	1	0.5	0.6667
Simple	Stacker	0.9412	0.6667	0.6667	0.6667

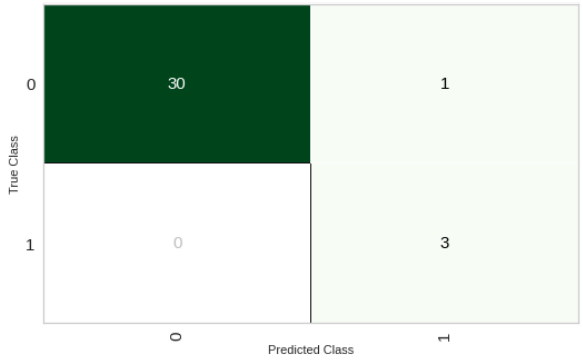

FIGURE 8. Confusion matrix of the best model (i.e., Simple-XGB) for death prediction.

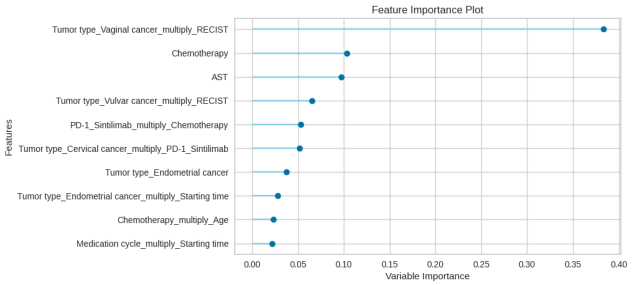

FIGURE 9. Feature importance of the best model (i.e., Simple-XGB) for death prediction.

d: FEATURE IMPORTANCE

Figure 9 shows the feature importance of Simple-XGB for death prediction. The importance scores are calculated by the random forest algorithm. Briefly, each DT in a random forest selects a feature to perform a split, and the feature that causes the most impurity decrease is selected and regarded as the most important one. It is observed that seven out of the top ten features are synthetic ones created by feature interaction, meaning that the interaction of features uncovers more informative patterns. The most important feature, Tumor type_Vaginal cancer_multiply_RECIST, indicates that a multiplication of vaginal cancer and RECIST score plays an important role in the prediction of death.

2) PREDICTION OF GENERAL AE

a: ABLATION STUDY

Figure 10 presents the box plots of the five-fold CV F1 scores of all fifteen models for the general AE prediction task

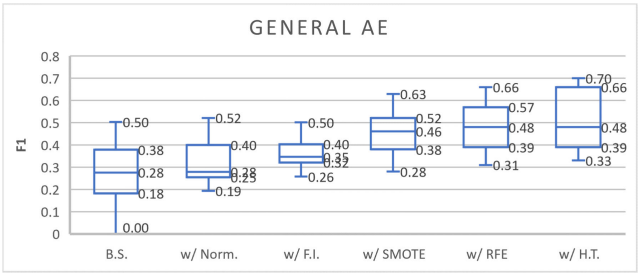

FIGURE 10. Results of general AE prediction under different training settings.

TABLE 7. Result of general AE prediction on the test set.

Imputation	Model	Accuracy	Recall	Precision	F1
Simple	LGB	0.6765	0.4545	0.5	0.4762
MICE	LR	0.5882	0.6364	0.4118	0.5
Iterative	SVM	0.7059	0.6364	0.5385	0.5833
KNN	KNN	0.6471	0.8182	0.4737	0.6
MissForest	NB	0.4706	**0.9091**	0.3704	0.5263
GAIN	KNN	0.7059	0.7273	0.5333	0.6154
GAIN	Blender	**0.7353**	0.7273	**0.5714**	**0.64**
GAIN	Stacker	0.5217	0.625	0.3846	0.4762

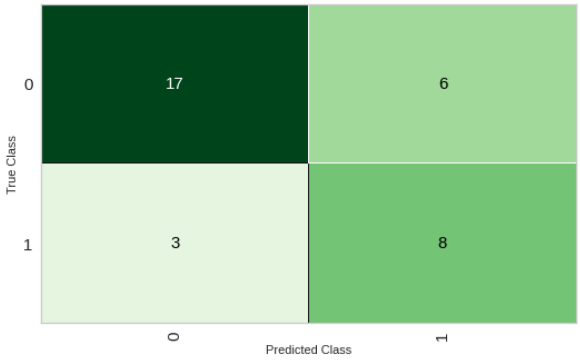

FIGURE 11. Confusion matrix of the best model (i.e., GAIN-Blender) for general AE prediction.

under different training settings. As more learning strategies applied, we can observe a consistent uptrend, with the best F1 rising from 0.5 in the base setting to 0.7 in the setting with all strategies applied.

b: OVERALL PERFORMANCE

As shown in Table 7, the best model for general AE prediction is GAIN-Blender, posting an accuracy of 0.7353, a recall of 0.7273, a precision of 0.5714, and an F1 of 0.64 (highest of all). Another model, MissForest-NB, presented the highest recall (0.9091) and the lowest precision (0.3704), indicating an unacceptable amount of false alarms, which makes the model useless.

c: CONFUSION MATRIX

Figure 11 shows the confusion matrix of the GAIN-Blender model for general AE prediction. The test set contains 23 negative cases and 11 positive cases, and our model misclassified six negative cases and missed three positive cases.

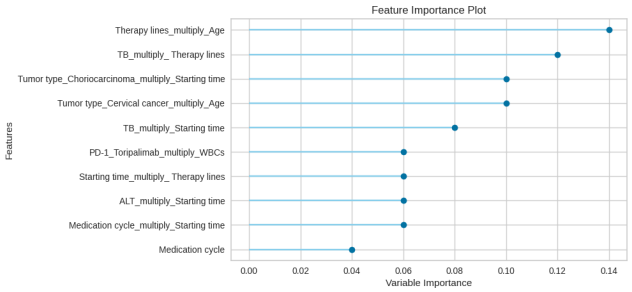

FIGURE 12. Feature importance of the best model (i.e., GAIN-Blender) for general AE prediction.

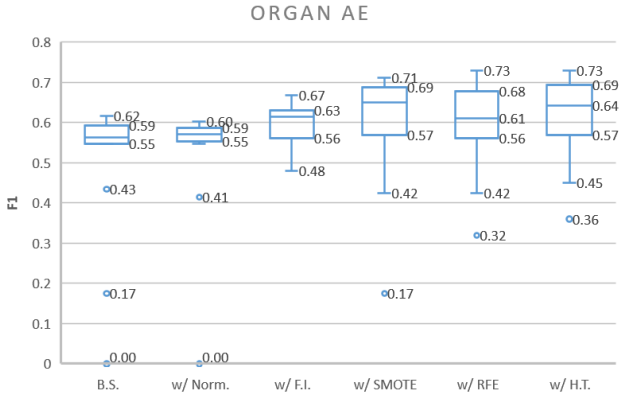

FIGURE 13. Results of organ AE prediction under different training settings.

d: FEATURE IMPORTANCE

Figure 12 shows the feature importance of GAIN-Blender for general AE prediction. It is observed that the top nine critical features are generated from feature interaction. A multiplication of therapy lines and age tops other features. The only non-synthetic feature is medication cycle, which is ranked the 10th place.

3) PREDICTION OF ORGAN AE

a: ABLATION STUDY

Figure 13 presents the box plots of the five-fold CV F1 scores of all fifteen models for the organ AE prediction task under different training settings. We can find similar effects on the model performance as more learning strategies are added to the base setting. The best model's F1 is elevated from 0.62 under the base setting to 0.73 when all strategies are adopted.

b: OVERALL PERFORMANCE

For the organ AE prediction task (Table 8), the best model is KNN-CB, which posted the highest accuracy (0.6176), recall (0.7222), and F1 (0.6667), with its precision (0.619) being the fourth highest of all models.

c: CONFUSION MATRIX

Figure 14 shows the confusion matrix of the KNN-CB model for organ AE prediction. The test set contains 16 negative

TABLE 8. Result of organ AE prediction on the test set.

Imputation	Model	Accuracy	Recall	Precision	F1
Simple	ADA	**0.6176**	0.6667	0.6316	0.6486
MICE	ADA	0.5882	0.5	**0.6429**	0.5625
Iterative	LR	0.5588	0.5556	0.5882	0.5714
KNN	CB	**0.6176**	**0.7222**	0.619	**0.6667**
MissForest	LDA	0.5882	0.5556	0.625	0.5882
GAIN	SVM	0.5588	**0.7222**	0.5652	0.6341
KNN	Blender	0.4706	0.5556	0.5	0.5263
KNN	Stacker	0.5	0.5556	0.5263	0.5405

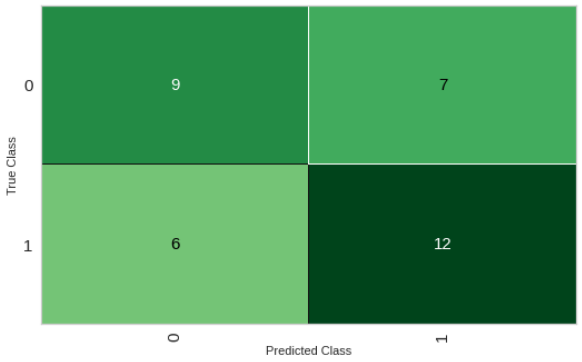

FIGURE 14. Confusion matrix of the best model (i.e., KNN-CB) for organ AE prediction.

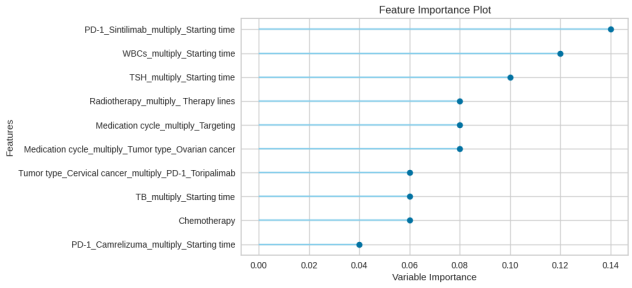

FIGURE 15. Feature importance of the best model (i.e., KNN-CB) for organ AE prediction.

cases and 18 positive cases, which is rather balanced. The KNN-CB model misclassified seven negative cases and missed six positive cases.

d: FEATURE IMPORTANCE

Figure 15 shows the feature importance of KNN-CB for organ AE prediction. We observe that most highly ranked features are synthetic via feature interaction. The starting time seems to take a crucial position for this task, since its interaction with PD-1 Sintilimab, WBCs, and TSH appear in the first three places among the features. The only non-synthetic feature in the top ten is Chemotherapy, ranked 9th place.

4) PREDICTION OF RECIST

a: ABLATION STUDY

We report the results of RECIST and binary RECIST prediction under different training settings in Figures 16 and 17. A consistent uptrend can be identified in both figures. For

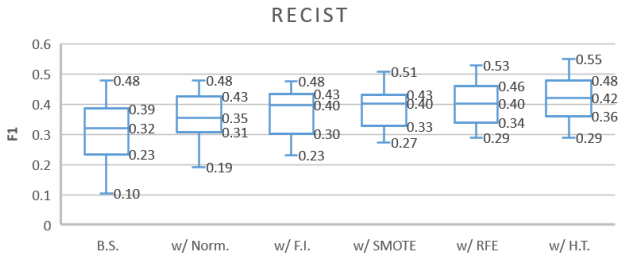

FIGURE 16. Results of RECIST prediction under different training settings.

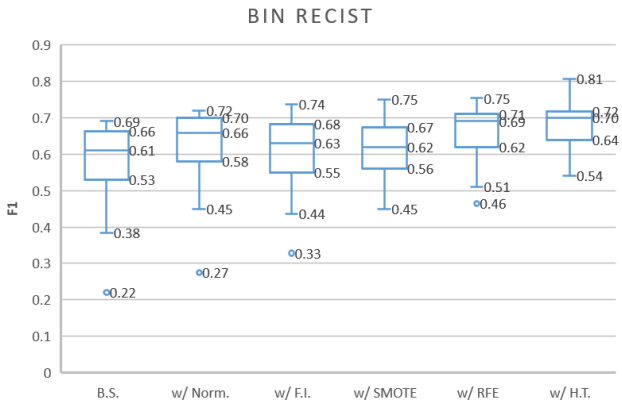

FIGURE 17. Results of binary RECIST prediction under different training settings.

TABLE 9. Result of RECIST prediction on the test set.

Imputation	Model	Accuracy	Recall	Precision	F1
Simple	ET	0.4229	0.3724	0.4301	0.4152
MICE	LR	**0.5043**	**0.4651**	**0.5442**	**0.5093**
Iterative	ET	0.4501	0.4026	0.4526	0.4406
KNN	ET	0.459	0.4163	0.4545	0.4492
MissForest	RF	0.5235	0.4517	0.5048	0.5019
GAIN	ET	0.4858	0.4361	0.4907	0.4734
MICE	blender	0.4286	0.3764	0.4732	0.4451
MICE	stacker	0.4233	0.3731	0.4612	0.4228

RECIST prediction, a 7-point gain on the best model is posted when moving from the base setting to the full setting. For binary RECIST, a 12-point gain on the best model is observed by switching from the base to full setting.

b: OVERALL PERFORMANCE

For the RECIST prediction task (Table 9), the best model was MICE-LR, which achieved the top performance for all four metrics. However, the performance scores were only in the range of 0.45 and 0.55, indicating a relatively low prediction ability. This is partly due to the hardness brought by multi-class classification, which would require more training data to build an accurate model. The best model for the binary RECIST prediction task (Table 10) was KNN-Stacker, which posted top performance in accuracy (0.7059), precision (0.8947), and F1 (0.7727).

TABLE 10. Result of binary RECIST prediction on the test set.

Imputation	Model	Accuracy	Recall	Precision	F1
Simple	RF	0.5588	0.64	0.7273	0.6809
MICE	ADA	0.6471	**0.72**	0.7826	0.75
Iterative	RF	0.5	0.56	0.7	0.6222
KNN	LR	0.6765	0.68	0.85	0.7556
MissForest	GBC	0.5588	0.64	0.7273	0.6809
GAIN	RF	0.5882	0.64	0.7619	0.6957
KNN	Blender	0.6471	0.64	0.8421	0.7273
KNN	Stacker	**0.7059**	0.68	**0.8947**	**0.7727**

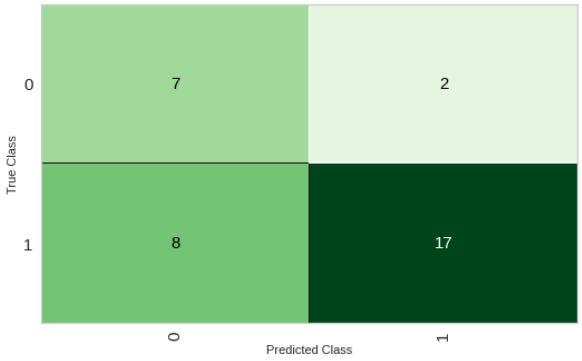

FIGURE 18. Confusion matrix of the best model (i.e., KNN-Stacker) for binary RECIST prediction.

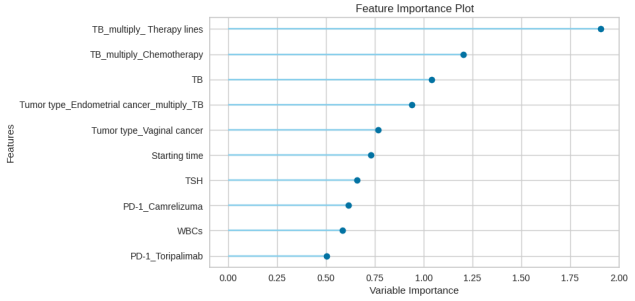

FIGURE 19. Feature importance of the best model (i.e., KNN-Stacker) for binary RECIST prediction.

c: CONFUSION MATRIX

Figure 14 shows the confusion matrix of the KNN-Stacker model for binary RECIST prediction. The test set contains nine negative cases and 25 positive cases. The KNN-Stacker model misclassified two negative cases and missed eight positive cases.

d: FEATURE IMPORTANCE

Figure 19 shows the feature importance of KNN-Stacker for binary RECIST prediction. Different from the other tasks, there are only three synthetic features among the top ten ranked ones. However, the top two features are synthetic. Also, feature TB plays a crucial role in this task as it appears in the top four features, either as a component in a synthetic feature or individually. Its role and function in the prediction of binary RECIST is worth further investigation.

V. DISCUSSION

ML has been a revolutionary technology that creates an explosive impact on all industries. Recent advances have explored the potential of ML models on outcome prediction of the anti-PD-1 therapy [7], [13]–[19]. Wiesweg *et al.* [18] adopt learning models to predict the response to anti-PD-1 therapy of non-small cell lung cancer (NSCLC). Arbour *et al.* develop a deep-learning model trained on radiology text reports to predict the gold-standard RECIST category [19]. Lewinson *et al.* train an artificial neural network model to estimate cutaneous adverse events for patients with metastatic melanoma or NSCLC [20]. Similar studies in the field of gynecological cancer have not been seen in the literature.

Inspired by the prior studies, we developed an analytical framework powered by ML to perform predictive analysis on the outcomes of anti-PD-1 therapy for patients with gynecological cancers. The study was challenged by the quality of a self-gathered post-marketing surveillance dataset, which is small, incomplete, and highly imbalanced. We employed data imputation methods to complete the data and the SMOTE technique to fix the imbalance problem. We investigated four therapeutic outcomes: RECIST, organ AE, general AE, and death, which were the four prediction targets. The dataset went through a learning pipeline consisting of imputation, feature engineering, base model training, ensemble learning, and model selection. The systematic process allows us to identify the most accurate and robust predictive model. In addition, we conducted extensive experiments to validate the proposed approach. The results can justify the design choices, such as the fold of cross-validation, the specific feature engineering strategies, and the SMOTE imbalance fixing technique. Our final model was selected from a large pool of model candidates, based on a joint consideration of F1 and accuracy, given the imbalanced nature of the dataset. Moreover, we conducted thorough and visualized model analysis and gained a deeper understanding of model behavior and feature importance.

The proposed analytical framework can be widely applied in clinical practice. A typical scenario is to assist practitioners in making therapy plans. Since the prediction result output by a model is a confidence score between 0 and 1, which can be interpreted as a probability of the outcome occurrence, a doctor can adjust the specific feature values that are a part of a therapy plan to observe the confidence change given by the model. For example, if switching to a different PD-1 inhibitor or adding a therapy program could boost the confidence of the CR outcome of RECIST, the doctor could take those adjustments into account, combined with the expert opinion, to generate a better therapy plan.

This study is subject to some limitations that also point to future directions. First, more patient samples can be added to the dataset to encourage the learning algorithms to discover more distinguishable patterns. We plan to continuously gather more patient data for training to make our model more robust. Also, more features, such as the metastasis type of cancer

and the other comorbidity, can be utilized in the dataset. Second, the study employed the SMOTE for fixing class imbalance, which worked well. However, we expect to conduct a further study on the effect of other data augmentation methods. Third, another meaningful work is to explore and quantify the joint effect of features, and to understand the causal relationship between the outcomes and feature factors, individually and collectively. A promising methodology that could benefit this objective would be probabilistic graphical models.

ACKNOWLEDGMENT

The funding agency has no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments, which are essential to the improvement of the manuscript quality.

REFERENCES

[1] M. Takahashi, M. Fujita, N. Asai, M. Saki, and A. Mori, "Safety and effectiveness of istradefylline in patients with Parkinson's disease: Interim analysis of a post-marketing surveillance study in Japan," *Expert Opinion Pharmacotherapy*, vol. 19, no. 15, pp. 1635–1642, Oct. 2018.

[2] M. Yi, D. Jiao, H. Xu, Q. Liu, W. Zhao, X. Han, and K. Wu, "Biomarkers for predicting efficacy of PD-1/PD-L1 inhibitors," *Mol. Cancer*, vol. 17, no. 1, pp. 1–14, Dec. 2018.

[3] J. Luo, H. Rizvi, J. V. Egger, I. R. Preeshagul, J. D. Wolchok, and M. D. Hellmann, "Impact of PD-1 blockade on severity of COVID-19 in patients with lung cancers," *Cancer Discovery*, vol. 10, no. 8, pp. 1121–1128, Aug. 2020.

[4] S. L. Topalian, F. S. Hodi, J. R. Brahmer, S. N. Gettinger, D. C. Smith, D. F. McDermott, J. D. Powderly, R. D. Carvajal, J. A. Sosman, M. B. Atkins, and P. D. Leming, "Safety, activity, and immune correlates of anti–PD-1 antibody in cancer," *New England J. Med.*, vol. 366, no. 26, pp. 2443–2454, 2012.

[5] A. Muls, J. Andreyev, S. Lalondrelle, A. Taylor, C. Norton, and A. Hart, "Systematic review: The impact of cancer treatment on the gut and vaginal microbiome in women with a gynecological malignancy," *Int. J. Gynecol. Cancer*, vol. 27, no. 7, pp. 1550–1559, Sep. 2017.

[6] G. Funston, H. O'Flynn, N. A. J. Ryan, W. Hamilton, and E. J. Crosbie, "Recognizing gynecological cancer in primary care: Risk factors, red flags, and referrals," *Adv. Therapy*, vol. 35, no. 4, pp. 577–589, Apr. 2018.

[7] C. Garcia and K. L. Ring, "The role of PD-1 checkpoint inhibition in gynecologic malignancies," *Current Treatment Options Oncol.*, vol. 19, no. 12, pp. 1–14, 2018.

[8] Y. Meng, H. Liang, J. Hu, S. Liu, X. Hao, M. S. K. Wong, X. Li, and L. Hu, "PD-L1 expression correlates with tumor infiltrating lymphocytes and response to neoadjuvant chemotherapy in cervical cancer," *J. Cancer*, vol. 9, no. 16, p. 2938, 2018.

[9] L. M. de la Fuente, S. Westbom-Fremer, N. S. Arildsen, L. Hartman, S. Malander, P. Kannisto, A. Måsbäck, and I. Hedenfalk, "PD-1/PD-L1 expression and tumor-infiltrating lymphocytes are prognostically favorable in advanced high-grade serous ovarian carcinoma," *Virchows Archiv*, vol. 477, no. 1, pp. 83–91, Jul. 2020.

[10] I. Puzanov, A. Diab, K. Abdallah, C. Bingham, C. Brogdon, R. Dadu, L. Hamad, S. Kim, M. E. Lacouture, N. R. LeBoeuf, and D. Lenihan, "Managing toxicities associated with immune checkpoint inhibitors: Consensus recommendations from the society for immunotherapy of cancer (SITC) toxicity management working group," *J. ImmunoTherapy Cancer*, vol. 5, no. 1, pp. 1–28, Dec. 2017.

[11] J. Naidoo, D. B. Page, B. T. Li, L. C. Connell, K. Schindler, M. E. Lacouture, M. A. Postow, and J. D. Wolchok, "Toxicities of the anti-PD-1 and anti-PD-L1 immune checkpoint antibodies," *Ann. Oncol.*, vol. 26, no. 12, pp. 2375–2391, Dec. 2015.

[12] M. A. Postow, R. Sidlow, and M. D. Hellmann, "Immune-related adverse events associated with immune checkpoint blockade," *New England J. Med.*, vol. 378, no. 2, pp. 158–168, 2018.

[13] Y. Muroyama, S. Manne, A. Huang, D. Mathew, L. Chilukuri, A. Greenplate, T. Ohtani, D. Zamarin, C. Friedman, and J. Wherry, "186 Distinct immune signatures predicting clinical response to PD-1 blockade therapy in gynecological cancers revealed by high-dimensional immune profiling," *J. ImmunoTherapy Cancer*, vol. 8, 2020, doi: 10.1136/jitc-2020-SITC2020.0186.

[14] T. J. Herzog, D. Arguello, S. K. Reddy, and Z. Gatalica, "PD-1, PD-L1 expression in 1599 gynecological cancers: Implications for immunotherapy," *Gynecol. Oncol.*, vol. 137, pp. 204–205, Apr. 2015.

[15] C. C. B. Post, A. M. Westermann, T. Bosse, C. L. Creutzberg, and J. R. Kroep, "PARP and PD-1/PD-L1 checkpoint inhibition in recurrent or metastatic endometrial cancer," *Crit. Rev. Oncol./Hematol.*, vol. 152, Aug. 2020, Art. no. 102973.

[16] O. Kooshkaki, A. Derakhshani, H. Safarpour, S. Najafi, P. Vahedi, O. Brunetti, M. Torabi, P. Lotfinejad, A. V. Paradiso, V. Racanelli, N. Silvestris, and B. Baradaran, "The latest findings of PD-1/PD-L1 inhibitor application in gynecologic cancers," *Int. J. Mol. Sci.*, vol. 21, no. 14, p. 5034, Jul. 2020.

[17] P. Song, X. Cui, L. Bai, X. Zhou, X. Zhu, J. Zhang, F. Jin, J. Zhao, C. Zhou, Y. Zhou, X. Zhang, K. Wang, Q. Wang, Y. Yu, X. Zhang, C. Bai, and L. Zhang, "Molecular characterization of clinical responses to PD-1/PD-L1 inhibitors in non-small cell lung cancer: Predictive value of multidimensional immunomarker detection for the efficacy of PD-1 inhibitors in Chinese patients," *Thoracic Cancer*, vol. 10, no. 5, pp. 1303–1309, May 2019.

[18] M. Wiesweg, F. Mairinger, H. Reis, M. Goetz, J. Kollmeier, D. Misch, S. Stephan-Falkenau, T. Mairinger, R. F. H. Walter, T. Hager, M. Metzenmacher, W. E. E. Eberhardt, G. Zaun, J. Köster, M. Stuschke, C. Aigner, K. Darwiche, K. W. Schmid, S. Rahmann, and M. Schuler, "Machine learning reveals a PD-L1–independent prediction of response to immunotherapy of non-small cell lung cancer by gene expression context," *Eur. J. Cancer*, vol. 140, pp. 76–85, Nov. 2020.

[19] K. C. Arbour, A. T. Luu, J. Luo, H. Rizvi, A. J. Plodkowski, M. Sakhi, K. B. Huang, S. R. Digumarthy, M. S. Ginsberg, J. Girshman, M. G. Kris, G. J. Riely, A. Yala, J. F. Gainor, R. Barzilay, and M. D. Hellmann, "Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade," *Cancer Discovery*, vol. 11, no. 1, pp. 59–67, Jan. 2021.

[20] R. T. Lewinson, D. E. Meyers, I. A. Vallerand, A. Suo, M. L. Dean, T. Cheng, D. G. Bebb, and D. G. Morris, "Machine learning for prediction of cutaneous adverse events in patients receiving anti–PD-1 immunotherapy," *J. Amer. Acad. Dermatol.*, vol. 84, no. 1, pp. 183–185, Jan. 2021.

[21] J. Zhang, Z. Wu, X. Zhang, S. Liu, J. Zhao, F. Yuan, Y. Shi, and B. Song, "Machine learning: An approach to preoperatively predict PD-1/PD-L1 expression and outcome in intrahepatic cholangiocarcinoma using MRI biomarkers," *ESMO Open*, vol. 5, no. 6, 2020, Art. no. e000910.

[22] G. S. Naik, "Anti-PD-1 based immunotherapy in melanoma: Application of machine learning to predict survival and elucidate complex relationships," M.S. thesis, Harvard Medical School, Harvard Univ., Boston, MA, USA, 2018.

[23] Y. Yan, S. Zeng, X. Wang, Z. Gong, and Z. Xu, "A machine learning algorithm for predicting therapeutic response to anti-PD1," *Technol. Cancer Res. Treatment*, vol. 18, Jan. 2019, Art. no. 153303381987576.

[24] J. Yang, W. Lin, L. Shi, M. Deng, and W. Yang, "A machine learning algorithm to predict hyperglycemic cases induced by PD-1/PD-L1 inhibitors in the real world," *Res. Square*, 2020.

[25] P. Johannet, N. Coudray, D. M. Donnelly, G. Jour, I. Illa-Bochaca, Y. Xia, D. B. Johnson, L. Wheless, J. R. Patrinely, S. Nomikou, D. L. Rimm, A. C. Pavlick, J. S. Weber, J. Zhong, A. Tsirigos, and I. Osman, "Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma," *Clin. Cancer Res.*, vol. 27, no. 1, pp. 131–140, Jan. 2021.

[26] B.-C. Ahn, J.-W. So, C.-B. Synn, T. H. Kim, J. H. Kim, Y. Byeon, Y. S. Kim, S. G. Heo, S. D. Yang, M. R. Yun, and S. Lim, "Clinical decision support algorithm based on machine learning to assess the clinical response to anti–programmed death-1 therapy in patients with non–small-cell lung cancer," *Eur. J. Cancer*, vol. 153, pp. 179–189, Aug. 2021.

[27] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[28] C. Dachena, S. Casu, A. Fanti, M. B. Lodi, and G. Mazzarella, "Combined use of MRI, fMRIand cognitive data for Alzheimer's disease: Preliminary results," *Appl. Sci.*, vol. 9, no. 15, p. 3156, Aug. 2019.

[29] K. K. Nagwanshi, "Learning classifier system," in *Modern Optimization Methods for Science, Engineering and Technology*. Bristol, U.K.: IOP, 2019, pp. 1–8 and 8–30, doi: 10.1088/978-0-7503-2404-5ch8.

[30] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 105–115, 2010.

[31] R. M. Thomas, W. Bruin, P. Zhutovsky, and G. van Wingen, "Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 249–266.

[32] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

[33] C. Wang and Z. Xiao, "Lychee surface defect detection based on deep convolutional neural networks with GAN-based data augmentation," *Agronomy*, vol. 11, no. 8, p. 1500, Jul. 2021. [Online]. Available: https://www.mdpi.com/2073-4395/11/8/1500

[34] S. Afzal, M. Maqsood, F. Nazir, U. Khan, and F. Aadil, "A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection," *IEEE Access*, vol. 7, pp. 115528–115539, 2019.

[35] S. G. K. Patro and K. K. Sahu, "Normalization: A preprocessing stage," 2015, *arXiv:1503.06462*. [Online]. Available: http://arxiv.org/abs/1503.06462

[36] B. Yang, S. Sun, J. Li, X. Lin, and Y. Tian, "Traffic flow prediction using LSTM with feature enhancement," *Neurocomputing*, vol. 332, pp. 320–327, Mar. 2019.

[37] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

[38] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.

[39] S. Basheer, K. K. Nagwanshi, S. Bhatia, S. Dubey, and G. R. Sinha, "FESD: An approach for biometric human footprint matching using fuzzy ensemble learning," *IEEE Access*, vol. 9, pp. 26641–26663, 2021.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[41] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemometric Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006.

[42] A. Markham and S. J. Keam, "Camrelizumab: First global approval," *Drugs*, vol. 79, no. 12, pp. 1355–1361, Aug. 2019.

[43] S. M. Hoy, "Sintilimab: First global approval," *Drugs*, vol. 79, no. 3, pp. 341–346, Feb. 2019.

[44] S. J. Keam, "Toripalimab: First global approval," *Drugs*, vol. 79, no. 5, pp. 573–578, Apr. 2019.

[45] J. P. T. Higgins, I. R. White, and A. M. Wood, "Imputation methods for missing outcome data in meta-analysis of clinical trials," *Clin. Trials*, vol. 5, no. 3, pp. 225–239, 2008.

[46] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, p. 9, 2016.

[47] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[48] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, Nov. 2012.

[49] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statist. Med.*, vol. 30, no. 4, pp. 377–399, Feb. 2011.

[50] D. J. Stekhoven and P. Bühlmann, "MissForest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012.

[51] J. Yoon, J. Jordon, and M. Schaar, "Gain: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[52] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *Int. J. Comput. Appl.*, vol. 175, no. 4, pp. 7–9, 2017.

[53] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.

[54] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Newton, MA, USA: O'Reilly Media, 2018.

[55] J. Su and H. Zhang, "A fast decision tree learning algorithm," in *Proc. AAAI*, vol. 6, 2006, pp. 500–505.

[56] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[57] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[58] D. Ruppert, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Oxfordshire, U.K.: Taylor & Francis, 2004.

[59] B. Efron and T. Hastie, *Computer Age Statistical Inference*, vol. 5. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[60] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[61] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.

[62] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[63] L. Alaskar, M. Crane, and M. Alduailij, "Employee turnover prediction using machine learning," in *Proc. Int. Conf. Comput.* Cham, Switzerland: Springer, 2019, pp. 301–316.

[64] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: An interdisciplinary review," *J. Big Data*, vol. 7, no. 1, pp. 1–45, 2020.

[65] S. Peter, F. Diego, F. A. Hamprecht, and B. Nadler, "Cost efficient gradient boosting," in *Proc. NIPS*, 2017, pp. 1551–1561.

[66] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[67] K. Fukunaga and P. M. Narendra, "A branch and bound algorithm for computing k-nearest neighbors," *IEEE Trans. Comput.*, vol. C-24, no. 7, pp. 750–753, Jul. 1975.

[68] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: A tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, pp. 145–180, 2016.

[69] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. Workshop Empirical Methods Artif. Intell. (IJCAI)*, vol. 3, no. 22, 2001, pp. 41–46.

[70] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc., B, Methodol.*, vol. 20, no. 2, pp. 215–232, 1958.

[71] S. L. Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 41, no. 1, pp. 191–201, 1992.

[72] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[73] S. Raschka, *Python Machine Learning*. Birmingham, U.K.: Packt, 2015.

[74] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," *Inf. Sci.*, vol. 384, pp. 174–190, Apr. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025516308957

[75] B. S. Raghuwanshi and S. Shukla, "SMOTE based class-specific extreme learning machine for imbalanced learning," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104814.

[76] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," in *Proc. 21st Australas. Comput. Sci. Conf. (ACSC)*, Perth, WA, Australia. Berlin, Germany: Springer, Feb. 1998, pp. 181–191.

[77] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1994, pp. 171–182.

[78] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013.

[79] M. Claesen and B. D. Moor, "Hyperparameter search in machine learning," 2015, *arXiv:1502.02127*. [Online]. Available: http://arxiv.org/abs/1502.02127

[80] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 1–25, 2012.

[81] T. Wu, W. Zhang, X. Jiao, W. Guo, and Y. A. Hamoud, "Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration," *Comput. Electron. Agricult.*, vol. 184, May 2021, Art. no. 106039.

[82] F. Divina, A. Gilson, F. Goméz-Vela, M. G. Torres, and J. F. Torres, "Stacking ensemble learning for short-term electricity consumption forecasting," *Energies*, vol. 11, no. 4, p. 949, Apr. 2018.
[83] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation metrics for evaluating classification performance on imbalanced data," in *Proc. Int. Conf. Comput., Control, Informat. Appl. (IC3INA)*, Oct. 2019, pp. 14–18.
[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

XIAOMEI LIU was born in 1983. She received the B.S. degree in clinical medicine from Jilin University, Changchun, Jilin, China, in 2015, and the M.S. degree in obstetrics and gynecology from Dalian Medical University, Changchun, in 2017. She is currently pursuing the Ph.D. degree in mechanical engineering with China Medical University, Shenyang, Liaoning, China. She is currently working as an Attending Physician with the Gynecologic Oncology Ward, Sheng Jing Hospital, China Medical University. Her research interests include treating gynecological malignant tumors, including comprehensive treatment, such as radiotherapy and chemotherapy, palliative treatment, and targeted immunotherapy.

ZHIFENG XIAO received the B.S. degree in computer science from Shandong University, China, in 2008, and the Ph.D. degree in computer science from The University of Alabama, in 2013. He is currently an Associate Professor with the Department of Computer Science and Software Engineering, Penn State Erie, The Behrend College. His research interests include interdisciplinary AI and cybersecurity, with a particular focus on AI-powered decision science, accountable systems, and bioinformatics.

YANG SONG was born in 1988. She received the B.S. degree in clinical medicine and the M.S. degree in obstetrics and gynecology from China Medical University, Shenyang, Liaoning, China, in 2011 and 2014, respectively, where she is currently pursuing the Ph.D. degree in mechanical engineering. She is currently working as an Attending Physician with the Gynecologic Oncology Ward, Sheng Jing Hospital, China Medical University. Her research interests include comprehensive treatment of gynecological tumors and gynecological endocrine therapy.

RUIZHE ZHANG received the B.S. degree in clinical medicine and the M.S. degree in oncology from Dalian Medical University, Dalian, Liaoning, China, in 2015 and 2017, respectively. His research interests include gynecology and oncology.

XIUQIN LI received the B.S. degree in clinical medicine and the M.S. degree in obstetrics and gynecology from China Medical University, Shenyang, Liaoning, China, in 1984 and 1993, respectively, where she is currently pursuing the Ph.D. degree in obstetrics and gynecology. She is currently the Chief Physician and a Professor of obstetrics and gynecology at Sheng Jing Hospital, China Medical University. She is engaged in comprehensive treatment and endocrine therapy of gynecological malignant tumors.

ZHENHUA DU received the B.S., M.S., and Ph.D. degrees in clinical medicine from China Medical University, Shenyang, Liaoning, China, in 2007, 2009, and 2019, respectively. She is currently an Associate Professor, an Associate Chief Physician, and a Supervisor for postgraduate students with the Department of Obstetrics and Gynecology, Sheng Jing Hospital, China Medical University. Her research interests include ovarian cancer anti-tumor immune microenvironment and recurrent cervical cancer.