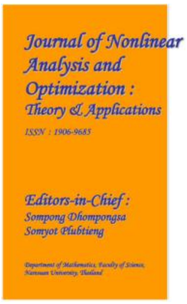
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.3 : 2024
ISSN : **1906-9685**



Journal of Nonlinear Analysis and Optimization : Theory & Applications
ISSN : 1906-9685

Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng

Department of Mathematics, Faculty of Science, Naresuan University, Thailand

ROLE OF MACHINE LEARNING IN CANCER DETECTION AND DIAGNOSIS: AN IMPLEMENTATION

Dr. Monalisa Hati, Assistant Professor, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharastra, India.
ssamit6@gmail.com
Sambhram Sampatkumar Jain, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharastra, India

Abstract:
It is studied that 8.8 million people died from cancer, making it the second most common cause of death worldwide. According to some descriptions, it is a diverse illness with several subgroups. Because it can help with the clinical care of patients later on, early detection and prognosis of a cancer type have become essential in cancer research. Accurately differentiating between benign and malignant tumors is crucial for better therapeutic decisions. Despite the intricate relationships between high-dimensional medical data, statistical techniques have historically been employed to classify cancer into high-risk and low-risk categories. Machine learning has become a promising tool for managing high-dimensional data, with increasing applicability in clinical decision support, to overcome the shortcomings of traditional statistical methods. This essay focuses on recent outlines research objectives and talks about the primary difficulties with machine learning techniques for cancer diagnosis.

Keywords: Cancer, Machine learning, Classification, Prediction and Diagnosis

1.INTRODUCTION

There are several linked disorders that share unchecked cellular growth and reproduction, including cancer. It kills almost 8 million people annually, making it the second largest cause of death in the poor world and the top cause in the developed world [1]. Because it can help with the clinical care of patients later on, early detection and prognosis of a cancer type have become essential in cancer research. Accurately differentiating between benign and malignant tumors is crucial for improved clinical judgment [2]. Despite the intricate relationships between high-dimensional medical data, statistical techniques have historically been employed to classify cancer into high-risk and low-risk categories [3]. In recent times, machine learning has been utilized to address the shortcomings of traditional statistical techniques. to the prediction and prognosis of cancer [4]. A subfield of artificial intelligence known as "machine learning" uses a range of statistical, probabilistic, and optimization methods to enable computers to "learn" from historical instances and identify challenging patterns in big, noisy, or complicated data sets [5]. Medical applications, particularly those that rely on intricate proteomic and genomic data, are especially well-suited to this capabilities. Consequently, machine learning is widely employed in the diagnosis and detection of cancer. This latter strategy is especially intriguing as it contributes to the expanding trend of predictive, tailored therapy [6]. Several tendencies are employed, such as an increasing reliance on microarray data and protein biomarkers, a strong bias towards applications in breast and prostate cancer, and a significant reliance on "older" technology.
using artificial neural networks (ANNs) in place of machine learning techniques that are more modern or simpler to understand. Additionally, some published research don't seem to have undergone enough testing or validation [7]. It is evident from the better-designed and verified research that machine learning techniques may significantly (15–25%) increase the accuracy of forecasting cancer mortality, recurrence, and susceptibility. More fundamentally, it is also clear that machine learning is contributing to a better understanding of the onset and course of cancer [8]. This study examines the

various machine learning techniques being applied, the kinds of data being combined, and how well these techniques detect and prognosticate cancer.

DIFFICULTIES IN CANCER PATIENT CLASSIFICATION

The availability of an adequate data base in terms of both number and quality is crucial to the success of contemporary, evidence-based, and customized medical research. This frequently alludes to subjects like data consolidation and interchange. A number of organizational, legal, and technological issues arise in the area of conflict between research interests, institutional frameworks, and data privacy [9]. One of the primary responsibilities of information management in medical research is to address these issues.

The marginal circumstances, needs, and idiosyncrasies of processing research data in the context of medical research are highlighted in case studies related to cancer research [10]. It is clear from study findings that cancer disorders are more like to disease families with several subtypes, and that the anatomical classification of Tumors may appear deceptive, and a categorization based on the pathological alteration of cellular signaling pathways is more appropriate. This distinction is crucial because, for one patient, a certain treatment may be highly relevant and effective, but for other patients with the "same" disease.

it only has negative consequences and no influence on tumor control [11].

The quantity and caliber of accessible data become crucial for evidence-based medicine to have a solid statistical foundation. The more pertinent aspects there are, the more data is needed. A wide range of elements and information are available when looking at contemporary cancer research, and this number is continually growing [12, 13]. In order to achieve evidence-based customized medicine, it is thus necessary to deal with this heterogeneity and to create sizable study bases by exchanging and combining medical research data.

Using machine learning techniques might be one approach to accomplish this aim.

Categorization of Machine Learning Methods for Cancer Identification

Computational tools and procedures have become crucial in this context thanks to recent technologies like microarray and next-generation sequencing. In cell biology, the dense nonlinear interactions between functional modules must be taken into account for many significant difficulties. It is now well acknowledged that computer simulation is essential to comprehending cellular processes, and several simulation techniques have been developed that are helpful for researching certain subsystems [14]. The machine learning approach generates a program by running input and output on the computer (Fig. 1).

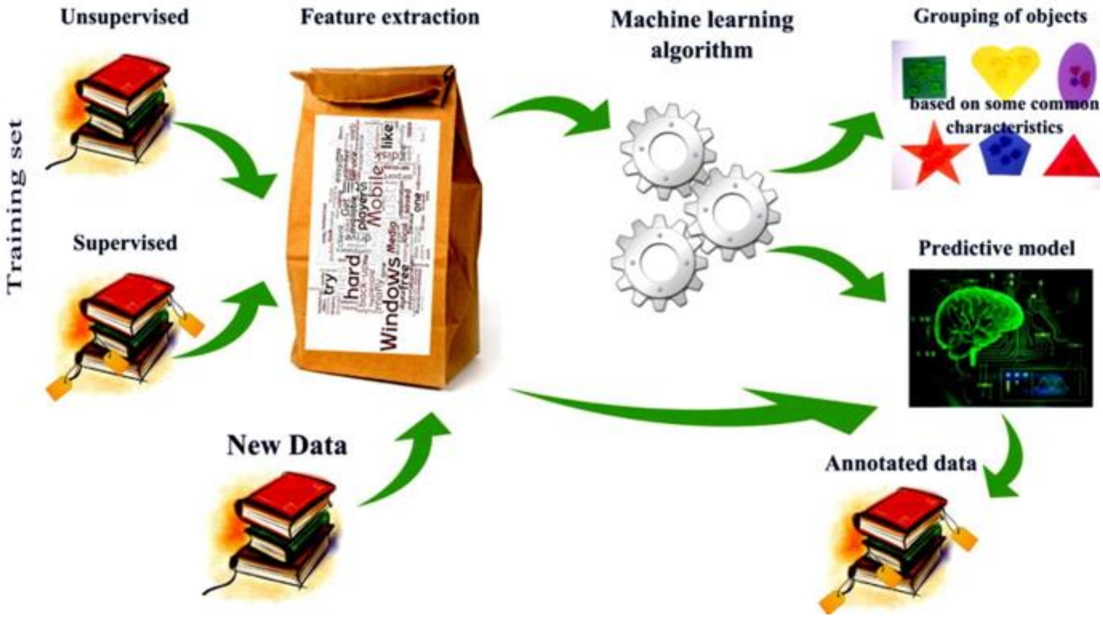


Figure: 1 Schematic representation of machine learning workflow
Source: www.google.com

1. Sparse compact incremental learning machine (SCILM) method

Using the correntropy cost, which makes it resilient to a variety of noises and outliers, a novel method was developed to address the issue of cancer classification using microarray gene expression data. Additionally, SCILM possesses sparseness since it employs the l1-norm of the weights, which may also be used for gene selection. Lastly, using just one neuron in its hidden layer, the approach can complete classification tasks in any scenario because of its compact nature. 26 well-known microarray data sets pertaining to various cancer types were used for the experimental analysis of this method. The findings demonstrated that the approach not only achieved a notably high accuracy but also that each gene was effective with regard to the corresponding cancer [15].

2. Knowledge base system learning method

Using clustering, noise reduction, and classification approaches, a novel knowledge-based system for classifying cancer diseases was put forth.To divide the data into comparable categories, Expectation Maximization (EM) was employed as a clustering technique. The fuzzy rules for the categorization of cancer illness in the knowledge-based system of fuzzy rule-based reasoning approach were created using categorization and Regression Trees (CART). Principal Component Analysis (PCA) was added to the existing knowledge-based system in order to address the multi-collinearity problem. The new method significantly increases the prediction accuracy of breast cancer, according on experimental results using the Wisconsin Diagnostic Breast Cancer and Mammographic Mass datasets.

Medical professionals can utilize the knowledge-based system as a clinical decision support tool to help them in their practice.

3.Gauss-Newton representation based learning method

A new method for classifying breast cancer that is based on Gauss-Newton representations (GNRBA). It selects training samples using the sparse representation. Sparse representation has only been effectively used in pattern recognition up to this point. In order to determine the ideal weights for the training samples for classification, this method presents a unique Gauss-Newton based strategy. Additionally, compared to the traditional l1-norm technique, it examines the sparsity in a computationally efficient manner.

The UCI Machine Learning repository's Wisconsin Breast Cancer Database (WBCD) and Wisconsin Diagnosis Breast Cancer (WDBC) databases are used to assess the GNRBA's efficacy.

4. Gene expression learning method

Analyzing hundreds of genes at once provided a comprehensive understanding of the cancer categorization issue. It brought in a wealth of data that was available for investigation. Additionally, it has been used in a variety of fields, including drug development, cancer diagnosis and prediction, and a crucial aspect of cancer therapy (Fig. 2). Additionally, it aids in comprehending how genes work and interact under both normal and aberrant circumstances [18]. This is accomplished by tracking gene expression data and gene behavior under various circumstances. Ensemble classifiers improve both the classification's performance and the findings' confidence. The results are less reliant on the quirks of a single training set, and the ensemble system performs better than the ensemble's top base classifier, which is another reason to employ ensemble classifiers [19].

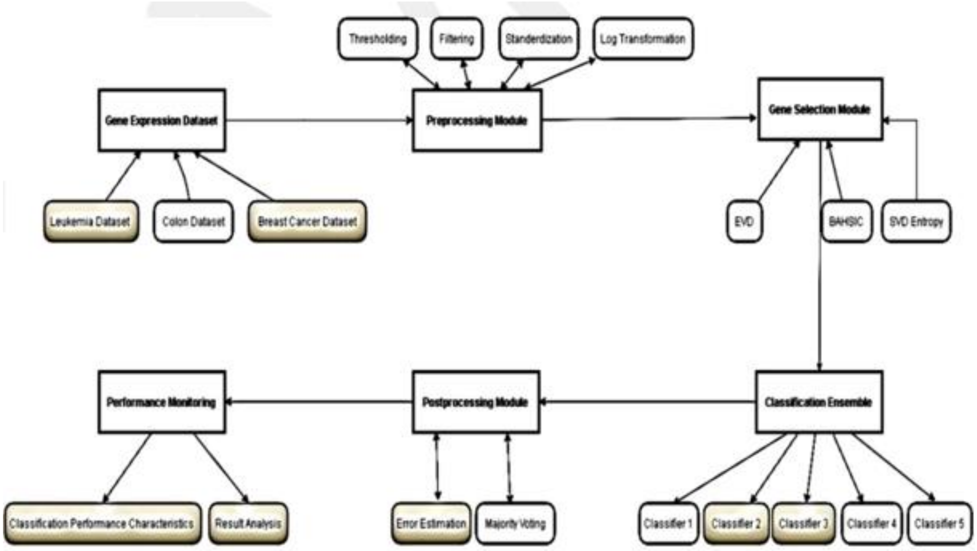
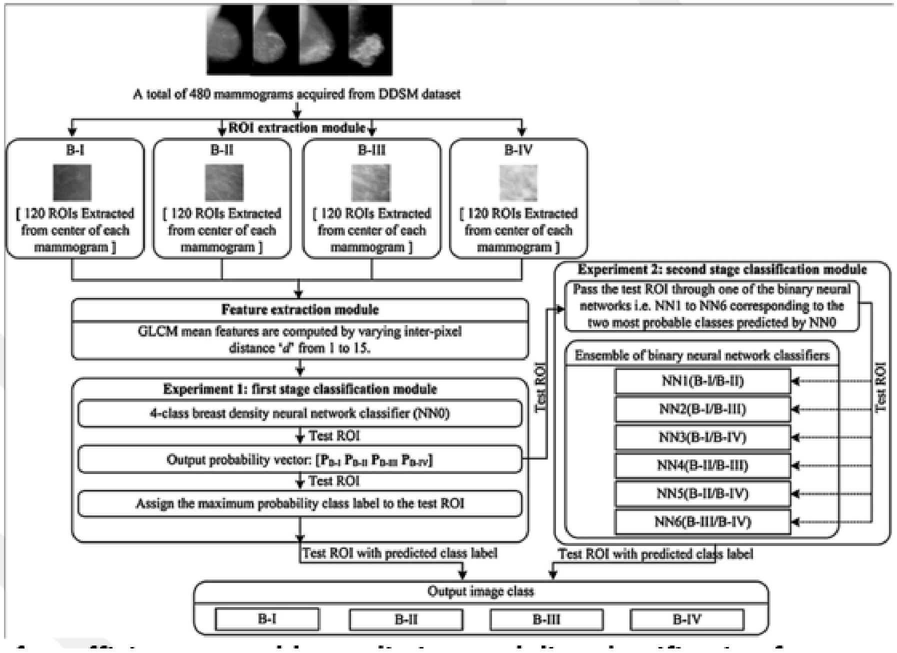


Figure: 2 Block diagram of gene expression based learning method [18].

5. Ensemble predictive modeling framework learning method

Molecular alterations frequently occur before a disease manifests clinically, and they can serve as helpful stand-ins that could support well-informed clinical judgment. The value of modeling techniques like classification that can forecast clinical outcomes using molecular expression patterns has been shown in recent research. Although helpful, most of these methods implicitly employ all molecular markers as features in the classification process, which frequently leads to a sparse high-dimensional projection of the samples that is frequently equivalent to the sample size. Based on their molecular expression patterns, breast cancer samples with favorable and bad prognoses are predicted using an ensemble classification technique [20]. The suggested method employs several base classifiers with different feature sets derived from two-dimensional projection of the data, as opposed to conventional single and ensemble classifiers. in addition to a majority vote method for class label prediction (Fig. 3). As opposed to previous implementations, basic classifiers in the ensembles are selected only based on low average cosine distance in order to maximize sensitivity and minimize redundancy. The ensemble sets that are produced are then represented as undirected graphs. Four distinct classification methods are demonstrated to perform better in the suggested ensemble framework than when used as conventional single classifier systems. [21].

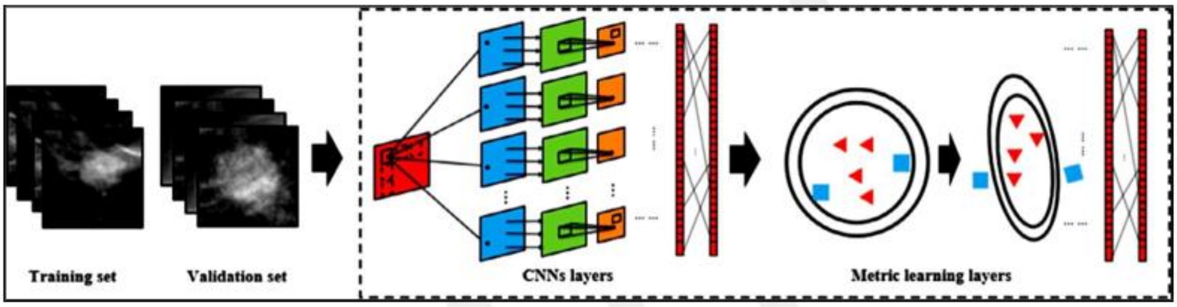

A total of 480 mammograms acquired from DDSM dataset
ROI extraction module
B-I
B-II
B-III
B-IV
[120 ROIs Extracted from center of each mammogram]
[120 ROIs Extracted from center of each mammogram]
[120 ROIs Extracted from center of each mammogram]
[120 ROIs Extracted from center of each mammogram]
Feature extraction module
GLCM mean features are computed by varying inter-pixel distance 'd' from 1 to 15.
Experiment 1: first stage classification module
4-class breast density neural network classifier (NN0)
Test ROI
Output probability vector: [P_B-I P_B-II P_B-III P_B-IV]
Test ROI
Assign the maximum probability class label to the test ROI
Test ROI
Experiment 2: second stage classification module
Pass the test ROI through one of the binary neural networks i.e. NN1 to NN6 corresponding to the two most probable classes predicted by NN0
Ensemble of binary neural network classifiers
NN1(B-I/B-II)
NN2(B-I/B-III)
NN3(B-I/B-IV)
NN4(B-II/B-III)
NN5(B-II/B-IV)
NN6(B-III/B-IV)
Test ROI
Test ROI with predicted class label
Test ROI with predicted class label
Output image class
B-I
B-II
B-III
B-IV

Source:www.google.com

Figure: 3 Design of an efficient ensemble predictive modeling classification framework for prediction of breast density class [21].

6. Convolution neural network learning method

The purpose of deep convolutional neural networks (CNNs) was to describe cancer tissues in a more discriminative manner. Metric learning layers are suggested to enhance the deep structure's performance even further by taking use of the discriminative representation. The top-performing model limits joint training's back propagation depth to only the metric learning layers. Metric learning layers and conventional CNN structures appear to have a parasitic relationship in which one species, the parasite, gains an advantage at the expense of the other [22]. Thus, the parasitic metric learning net is the term given to the technique. Classification studies on breast mass photos from two popular databases were conducted to verify the validity of this approach (Fig. 4). Competitive results were obtained when comparing the current method's performance to those of conventional methods. In the meantime, A method for enhancing the performance of a pre-trained CNN model on specific medical image processing or other computer vision applications may be inspired by the parasitic metric net parameter update technique [23].



Source: www.google.com

Figure: 4 Training pipeline of the proposed parasitic metric learning net [23].

2.METHODOLOGY

Description of the data.

Open Availability of TCMA20, located at https://tcma.pratt.duke.edu, was used to obtain the data. This database guarantees that the required ethical clearances for data access were acquired for the corresponding published descriptive paper. It offers 620 samples in total and allows users to select the taxonomy level of the microbiological data based on phylum, class, order, family, and genus. The microbiological data used in this study was gathered at the genus level, encompassing 221 distinct taxa.

The TCMA database provides information in two distinct but complementary datasets: the metadata dataset (which includes information on the type of sample, cancer type, and related TCGA project) and the microbial dataset.

This includes a unique sample and the relative abundance of the corresponding genus in a given sample).

Pre-processing

In order to exclude any information that was either insufficient or unrelated to this inquiry, the initial stage of data pre-processing was a general evaluation of the data sets. This characteristic can be used to indicate either the normal tissue next to the tumor normal (STN) or the tumor tissue of the main tumor (PT), depending on the kind of sample. 512 (82.58%) of the 620 samples were PT samples. The remaining 108 STN samples were eliminated since they were not relevant to the study's concept, which was to differentiate between different forms of cancer by analyzing their unique microbiological information. The sample distribution for the different malignancies was as follows, based on the kind of cancer and the associated TCGA project: HNSC (155 specimens), The distribution of the different malignancies was as follows: 155 samples of HNSC, 127 samples of STAD, 125 samples of COAD, 60 samples of ESCA, and 45 samples of READ.Additionally, it was noted that several of the 221 taxa that were originally included in the dataset were absent from all of the samples. Since they did not provide any useful information for the study, these non-present genera were not included in the analysis. The final dataset of 512 samples with the relative abundance of 131 genera utilized as features for the ML models was therefore obtained by removing 90 genera from the original set of 221 (see

Supplementary Table S1). Interestingly, the TCMA database data was already standardized, taking into account the percentage of each species relative to the total quantity of germs in the sample, removing the need for further normalizing procedures.

Design and measurements of the experiment.

It is crucial to provide the ML models the most informative characteristics available using techniques like feature selection because the microbiological dataset in TCMA is quite sparse. In this way, the RF algorithm, which has an inherent feature selection mechanism, served as the foundation for the ML method that was developed. The scikit-learn package21 was used to write code in the Python programming language. To guarantee consistency between trials, the RF models were trained and evaluated on distinct, stratified sample splits of 85% and 15% of the dataset, respectively. Grid search optimization with stratified 5-fold cross-validation on the training split was used to do hyper-parameter tweaking with the goal of maximizing the balanced accuracy of the validation set's RF model. To evaluate the prediction value of the microbiological data on different degrees of specificity depending on the anatomic location of the various cancer types, a total of four granularity levels of analysis were carried out. A one-vs-all approach to the classification problem was used in the first trial, which gave researchers a preliminary understanding of how well the RF model performed when distinguishing between individual cancers. The five cancer types included in the TCMA database—HNSC, STAD, COAD, ESCA, and READ—were then combined into three main groups according to anatomical closeness for a second study: HNSC, STAD/ESCA, and colorectal cancer (CRC)17,22,23 (Fig. 1a). This research provided a foundation for assessing the capacity of the microbiological information in the categorization of various anatomical regions where the cancer was found, and would also enable the research to move forward in a direction of more precision about the cancer location. An even more precise method was used in the third trial, where CRC was kept as a combination of COAD and READ while STAD and ESCA were divided into their original groups (Fig. 1b).

The five first classifications that TCMA offered were the result of splitting CRC into COAD and READ in the fourth and final investigation, which is the most fine-grained technique (Fig. 1c). Fig. 2 shows the experiment process for the learning model development for each research of the detection's granularity. It includes the tests listed below:

• Experiment 1: The algorithm's initial isolated implementation and performance evaluation following hyper-parameter tuning;

Experiment 2: A number of dimensionality reduction techniques, such as Sparse Principal Component Analysis (SPCA), Non-negative Matrix Factorization (NMF), and Linear Discriminant Analysis (LDA), were tested in an effort to give the models a simpler and more separable feature space and ultimately improve the performance of the baseline RF models.

Hyper-parameter tuning was done for the auxiliary approaches.

These were used separately in conjunction with the adjusted RF algorithm.

• Experiment 3: A follow-up experiment was created based on a feature in the event that the dataset's performance was not significantly enhanced by applying dimensionality reduction engineering approach where the components given by the dimensionality reduction techniques were added to the dataset while maintaining the original feature set;

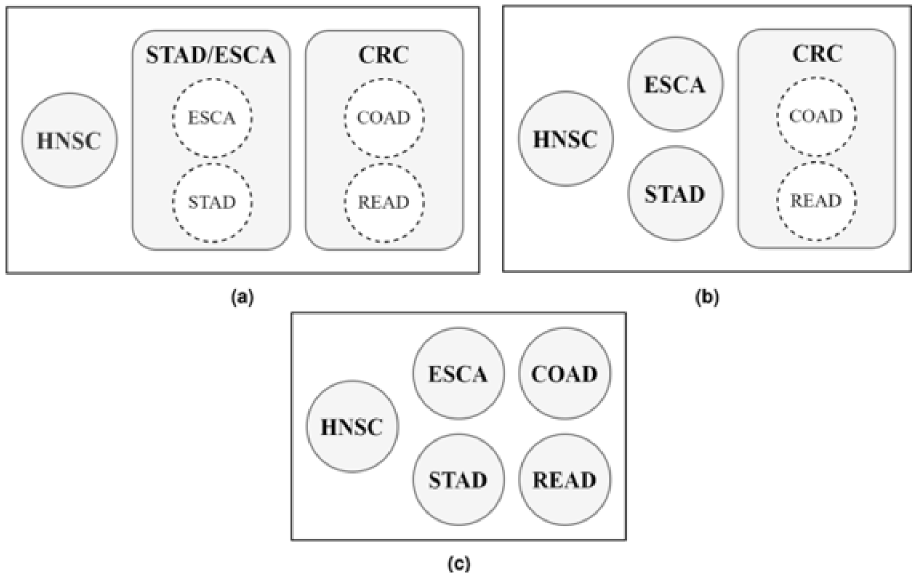


Figure 1. *Multi-class classification studies conducted with cancer types grouped according to **(a)** 3-class, **(b)** 4-class, and **(c)** 5-class approaches.*

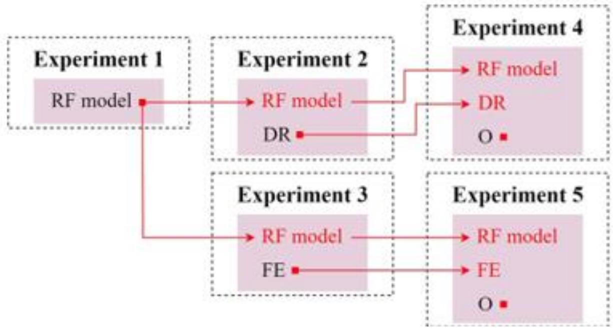


Figure 2. *Experiment pipeline for the learning models development, with the isolated implementation of the RF algorithm (Experiment 1), the implementation of dimensionality reduction (DR) and feature engineering (FE) alongside the tuned RF (Experiments 2 and 3, respectively), and the final application of oversampling (O) with the previous techniques (Experiments 4 and 5). Red squares indicate that hyper-parameter tuning is being performed to a specific technique while red arrows indicate the use of a technique tuned in a former experiment*

• Experiment 4: Steps must be taken to mitigate the potential negative effects of the current class imbalance, taking into account the variations in sample sizes for each class. In this regard, data augmentation was also utilized, with Random Oversampling and SVM-SMOTE being employed as oversampling approaches, with the primary goal of enhancing the performance of RF models when classifying samples from the minority classes. In this context, each model was built using dimensionality reduction and oversampling, concurrently. The dataset was first subjected to the dimensionality reduction technique that produced the greatest improvement in performance in Experiment 2, and then oversampling techniques were adjusted accordingly.

• Experiment 5: comparable to experiment 4, but feature engineering is used in place of dimensionality decrease. Lastly, it is crucial to take into account the varying sample sizes across classes when evaluating the ML model's performance metrics. Because balanced accuracy accounts for these differences and does not signal high performance when the model exploits the majority classes, it was selected as the metric to evaluate the performance of the ML models.

Confusion matrices were also created since they make it easier to comprehend how the model behaves while classifying samples for each distinct class.

Baseline is the first experiment. As a baseline model for the subsequent experiments pertaining to dimension reduction and feature engineering, the first experiment comprised an isolated implementation and hyper-parameter tuning of the RF algorithm.

Supplementary Table S2 lists the hyper-parameters and the range of values that are utilized for tuning.

Experiments 2 and 3: Feature engineering and dimensionality reduction. The hyper-parameters of the predictive model were fixed based on the baseline in order to tune the dimensionality reduction strategies. In addition to LDA, the number of components provided by SPCA and NMF varied from 4 to 64 in the feature engineering technique and from 10 to 100 in the dimensionality reduction strategy. In the instance of LDA, the number of components varied between 1 and n − 1 for both methods as its output must have a dimension smaller than a specified number of classes n.

Experiments 4 and 5: oversampling and dimensionality reduction/feature engineering. In light of the implementation of Oversampling was used in the training split and dimensionality reduction/feature engineering in the dataset with the goal of enhancing the models' performance, particularly in identifying samples that belong to the minority classes. Depending on the degree of oversampling, classes were oversampled using three different methods. Each class was oversampled in the first method until it had the same number of samples as the majority class, which would have kept the initial sample size.

All other classes matched the sample size of the majority class, which was oversampled by 50% and 100% in the second and third methods, respectively.

4. Findings

One-versus-all is the first study. Five separate analyses were carried out in the one-versus-all analysis. A binary classification challenge was created by targeting a separate cancer class (positive class) for each study and combining the data from the other classes into a single major class (negative class).

Table 1 provides a summary of the results, while Supplementary Tables S3 through S7 offer a more thorough analysis. The five tumors create two groups with different performances based on the test split's balanced accuracy of the RF models. Results for ESCA and READ indicate a rise in categorization complexity, with performances for both instances below, whereas HNSC, STAD, and COAD attained balanced accuracy rates ranging from 87% to 96%.

	Balanced accuracy
HNSC versus All	87.38 ± 2.19
STAD versus All	92.04 ± 1.02
COAD versus All	96.21 ± 0.42
ESCA versus All	72.35 ± 3.11
READ versus All	78.86 ± 6.15

Table 1. One-versus-All study: best performance in the test split of the RF model for each cancer type. Results in more detail are found in Supplementary Tables S3 to S7. Results from 5-fold cross-validation are given as the balanced accuracy of the model in mean (%) ± standard deviation (%) format.

80 percent. The related confusion matrices (Fig. 3) provide more information on these findings. The model accurately categorized all samples of this form of cancer, demonstrating that the microbial composition of COAD was the most discriminative.

However, the confusion matrix from the ESCA analysis shows that the low accuracy of 64% in classifying ESCA samples—a notable discrepancy in performance when compared to the results from the other cancer types—is the main cause of the RF model's poor balanced accuracy. The one-versus-all study's findings partially showed that microbiological data may be successfully used to independently and reliably categorize different cancer kinds. However, there are also significant differences in how well each type of cancer does.

S. While classes with smaller sample sizes tend to have inferior outcomes, they may also indicate different levels of complexity and the need to modify ML implementations and the microbiological data presented based on the kind of cancer. Additionally, the research by Poore et al.18 included a comparable analysis of a supervised machine learning model on a one-versus-all method that discriminated among other things these distinct cancer kinds.

However, the comparison of results was undermined by the wide number of distinct cancer types that were included in the research as well as the significant sample size discrepancy—in many cases, more than ten times the number of samples supplied by TCMA for a particular malignancy. Study 2: exam in three classes.

Following an evaluation of the RF's ability to distinguish between a certain kind of The remaining five kinds of cancer were divided into three primary categories:

HNSC, STAD/ESCA, and CRC (COAD and READ).

The RF model with oversampling and dimensionality reduction as auxiliary strategies produced the greatest balanced accuracy of 88% across all experiments, which was a 4% improvement over the baseline model's performance (Table 2).

With the RF maintaining accuracy levels over 90%, CRC seems to be the most readily separable of the three groups, which is in line with the findings of the one-versus-all investigation (Fig. 4).

The RF's performance in differentiating HNSC cancer cases was primarily enhanced by the dimensionality reduction and oversampling procedures; its classification success rate increased from 70% to 87%

Even though it was evident that the RF model's isolation performance was improved by applying dimensionality reduction and oversampling to the dataset, the benefits only applied to the HNSC class. The tests failed to improve performance in the STAD/ESCA class, since the categorization of CRC cancer patients already reaches high accuracy levels. Additional analyses were carried out using ESCA and STAD as separate classes in order to evaluate this scenario in greater detail.

Study 3: test with four classes. The prior study's results demonstrated the RF's capacity to accurately categorize the HNSC, CRC, and STAD/ESCA classes. In further detail, however, the model's performance in categorizing STAD/ESCA cancer data was not improved by dimensionality reduction and oversampling strategies.

increasing accuracy ratings only in relation to the HNSC class. STAD and ESCA samples were examined independently in order to have a better understanding of the prediction limits of the microbiological data on a more detailed level of specificity.

With feature engineering and oversampling improving the baseline result by 7%, the RF attained an overall performance of 74% balanced accuracy in this four-class classification task (Table 3). Oversampling was identified as a critical strategy to achieve the best performance from the model since the sample distribution for each class, as shown in Supplementary Figure S1, shows a significant imbalance between ESCA and the remaining classes. However, the confusion matrix clearly shows the notable decline in overall performance when compared to earlier research (Fig. 5). The RF displayed a only 50% of instances were accurately classified as ESCA samples, with the other cases being incorrectly predicted as HNSC or STAD (Fig. 5c).

ESCA is proving to be a significant constraint to the otherwise encouraging scores, as the model reveals a reasonable success rate of 75% in the categorization of STAD cancer samples. This notion was previously validated by the findings of the one-versus-all research.

The results clearly reveal a reduction in prediction ability from the microbiological data with a better level of specificity in terms of cancer location. The RF shows poorer performance in categorizing HNSC cancer patients compared to the three-class test, and it is particularly challenging to distinguish ESCA cases from STAD. However, With accuracy values exceeding 95%, the CRC class remains the most discriminative. The next investigation was carried out with both cancer kinds treated independently since the CRC class is made up of instances of COAD and READ.

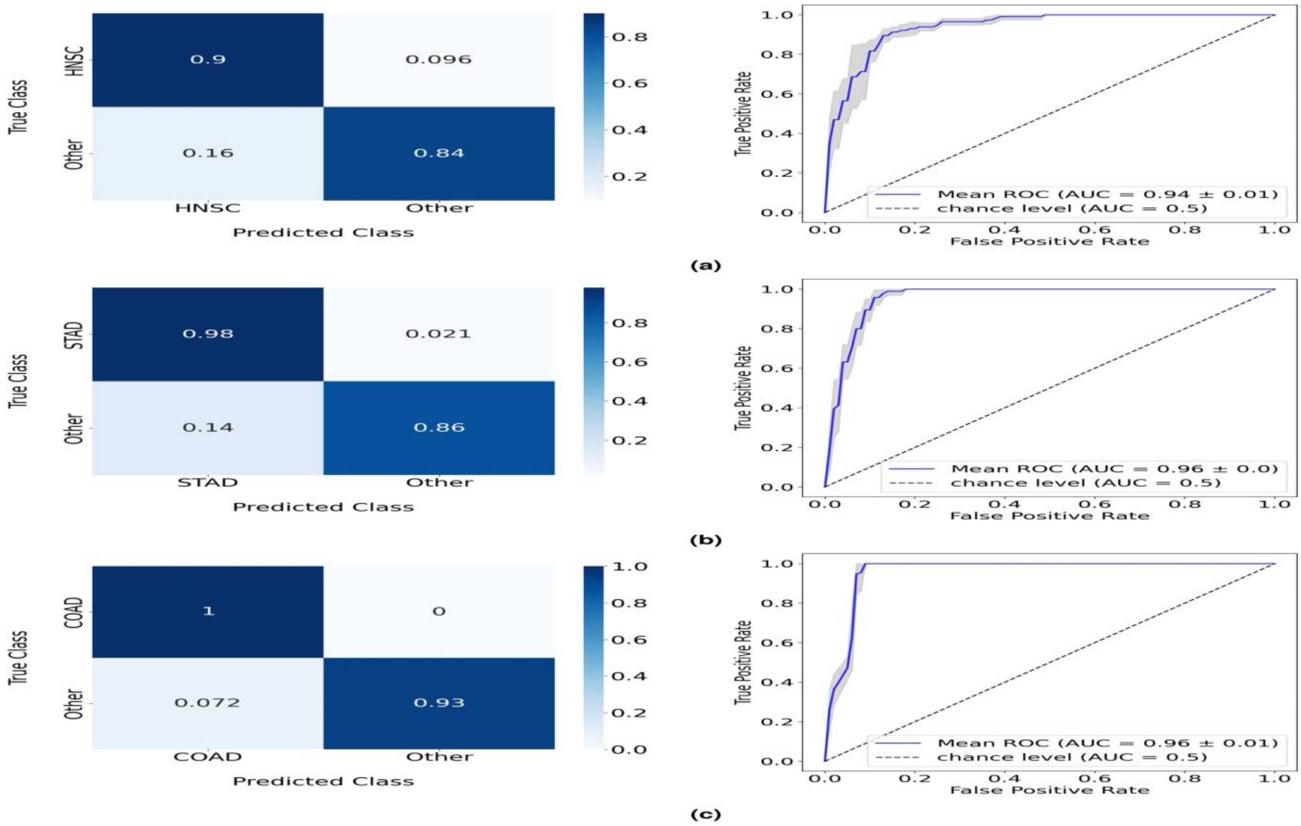
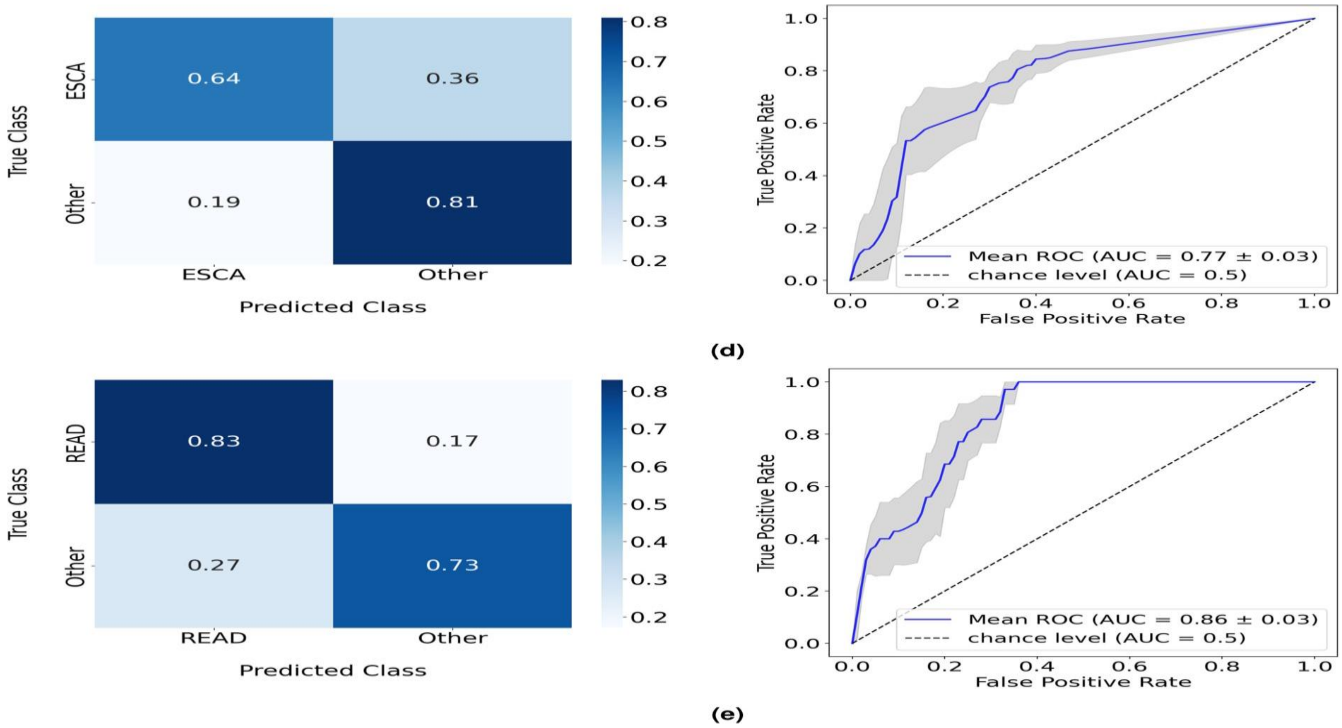


Figure 3. *Normalized confusion matrices and corresponding ROC curves of the RF performances in the one- versus-all study, targeting (**a**) HNSC, (**b**) STAD, (**c**) COAD, (**d**) ESCA, and (**e**) READ cancer cases.*



	Balanced accuracy
Experiment 1: RF	83.84 ± 3.06
Experiment 2: RF + Dimensionality reduction	86.13 ± 3.15
Experiment 3: RF + Feature engineering	86.25 ± 1.77
Experiment 4: RF + Dimensionality Reduction + Oversampling	88.28 ± 1.63
Experiment 5: RF + Feature engineering + oversampling	87.05 ± 2.30

Table 2. *Performance in the test split of the RF model in the three-class study. Results in more detail are found in Supplementary Table S8. Results from 5-fold cross-validation are given as the balanced accuracy of the model in mean (%) ± standard deviation (%) format.*

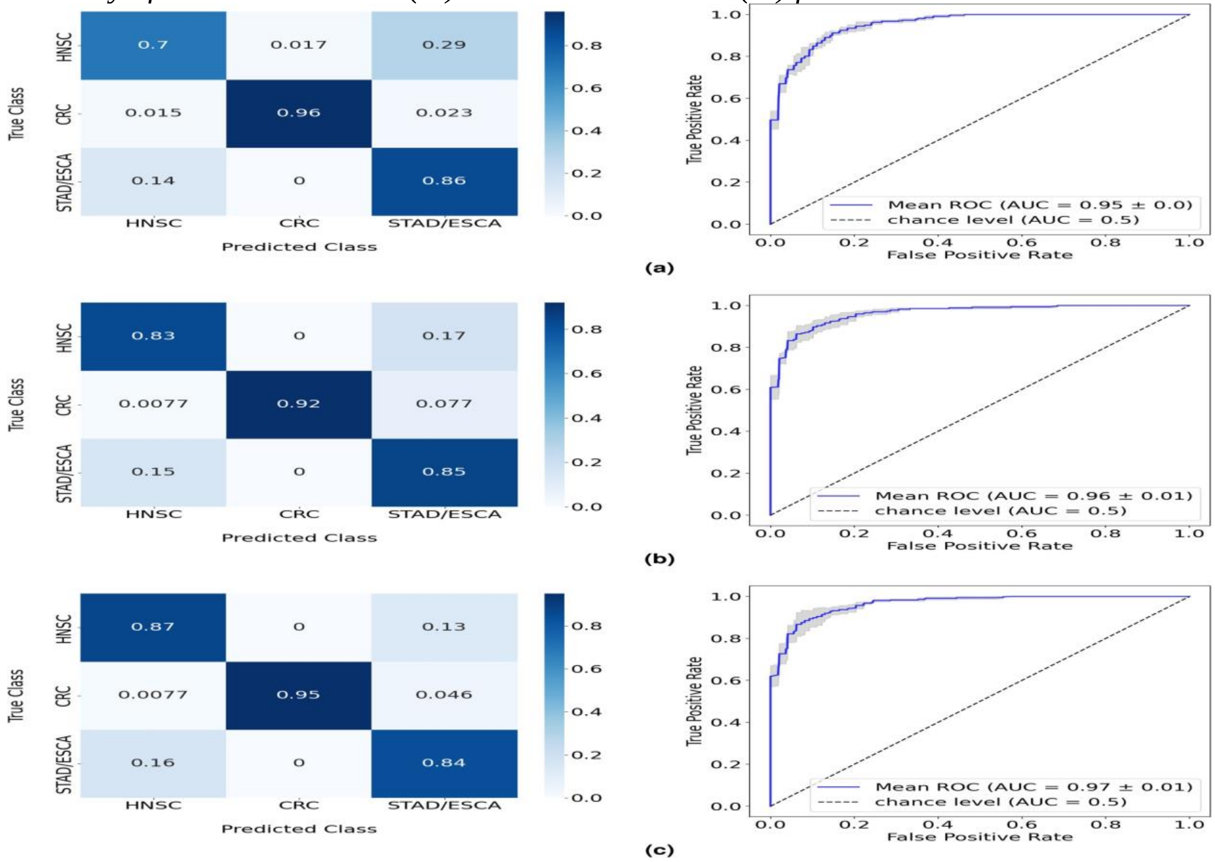


Figure 4. *Normalized confusion matrices and corresponding ROC curves of the RF performances in the three- class test, with (**a**) the isolated implementation of RF, (**b**) the implementation of dimensionality reduction, and*
*(**c**) the introduction of oversampling to the model alongside dimensionality reduction*

	Balanced accuracy
Experiment 1: RF	67.21 ± 4.50
Experiment 2: RF + Dimensionality reduction	68.86 ± 1.82
Experiment 3: RF + Feature engineering	70.11 ± 4.76
Experiment 4: RF + Dimensionality reduction + Oversampling	72.80 ± 4.91
Experiment 5: RF + Feature engineering + Oversampling	74.06 ± 4.53

Table 3. *Performance in the test split of the RF model in the four-class study. Results in more detail are found in Supplementary Table S9. Results from 5-fold cross-validation are given as the balanced accuracy of the model in mean (%) ± standard deviation (%) format.*

5.CONCLUSION

In this review, we went over machine learning techniques and how they are used in cancer diagnosis and prognosis. In order to anticipate accurate illness outcomes, the majority of research that have been proposed in recent years concentrate on the creation of predictive models employing supervised machine learning techniques and classification algorithms. It is clear from the examination of their findings that combining multidimensional heterogeneous data with various feature selection and classification methods might yield innovative inference tools for the cancer field.

This work's primary objective was to create a machine learning method for differentiating between cancer kinds by analyzing their unique microbiological data. The categorization of HNSC, STAD, COAD, ESCA, and READ malignancies was accomplished by training RF models for this purpose; dimensionality reduction and oversampling approaches helped to increase performance.

6.REFERENCES

1. G. D. Magoulas & A. Prentza (2001). Machine Learning in Medical Applications. Machine Learning and Its Applications Lecture Notes in Computer Science, 300- 307 Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).

2. de Martel, C., Georges, D., Bray, F., Ferlay, J. & Clifford, G. M. Global burden of cancer attributable to infections in 2018: A world- wide incidence analysis. *Lancet Glob. Health* **8**, e180–e190 (2020).

3. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).

4. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).

5. Ferreira, R. M. *et al.* Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. *Gut* **67**, 226–236 (2018).

6. Hieken, T. J. *et al.* The microbiome of aseptically collected human breast tissue in benign and malignant disease. *Sci. Rep.* **6**, 1–10 (2016).

7. Zheng, Y. *et al.* Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* **11**, 1030–1042 (2020).

8. Pereira-Marques, J., Ferreira, R. M., Pinto-Ribeiro, I. & Figueiredo, C. Helicobacter pylori infection, the gastric microbiome and gastric cancer. *Helicobacter pylori Hum. Dis.* **11**, 195–210 (2019).

9. Ma, C. *et al.* Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells. *Science* **360**, eaan5931 (2018).

10. Rodriguez, R. M., Hernandez, B. Y., Menor, M., Deng, Y. & Khadka, V. S. The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing. *Comput. Struct. Biotechnol. J.* **18**, 631–641 (2020).

11. Nejman, D. *et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973–980 (2020).

12. Narunsky-Haziza, L. *et al.* Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. *Cell* **185**, 3789–3806 (2022).

13. Dohlman, A. B. *et al.* A pan-cancer mycobiome analysis reveals fungal involvement in gastrointestinal and lung tumors. *Cell* **185**, 3807–3822 (2022).

14. Eisenhofer, R. *et al.* Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends Microbiol.* **27**, 105–117 (2019).

15. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 1–12 (2014).

16. Wu, H. *et al.* Metagenomics biomarkers selected for prediction of three different diseases in Chinese population. *BioMed Res. Int.*https://doi.org/10.1155/2018/2936257 *(2018).*

17. Baxter, N. T., Ruffin, M. T., Rogers, M. A. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemi- cal test for detecting colonic lesions. *Genome Med.* **8**, 1–10 (2016).

18. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

19. Dadkhah, E. *et al.* Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterol.* **6**, e000297 (2019).

20. Dohlman, A. B. *et al.* The cancer microbiome atlas: A pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* **29**, 281–298 (2021).

21.Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

22.Wu, Y. *et al.* Identification of microbial markers across populations in early detection of colorectal cancer. *Nat. Commun.* **12**, 1–13 (2021).

23.Sánchez-Alcoholado, L. *et al.* The role of the gut microbiome in colorectal cancer development and therapy response. *Cancers* **12**, 1406 (2020).

24.Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) (Curran Associates Inc., 2017).

25.da Costa, C. P. *et al.* The tissue-associated microbiota in colorectal cancer: A systematic review. *Cancers* **14**, 3385 (2022).

26.Wang, Y. *et al.* Analyses of potential driver and passenger bacteria in human colorectal cancer. *Cancer Manag. Res.* **12**, 11553 (2020).

27.Wang, X. *et al. Porphyromonas gingivalis* promotes colorectal carcinoma by activating the hematopoietic NLRP3 inflammasome- porphyromonas gingivalis promotes colorectal carcinoma. *Can. Res.* **81**, 2745–2759 (2021).

28.Mu, W. *et al.* Intracellular porphyromonas gingivalis promotes the proliferation of colorectal cancer cells via the MAPK/ERK signaling pathway. *Front. Cell. Infect. Microbiol.* **10**, 584798 (2020).

29.Huang, Y. *et al.* Is laryngeal squamous cell carcinoma related to *Helicobacter pylori*?. *Front. Oncol.* **12**, 790997 (2022).

30.Pandey, S. *et al.* Helicobacter pylori was not detected in oral squamous cell carcinomas from cohorts of Norwegian and Nepalese patients. *Sci. Rep.* **10**, 1–8 (2020).

31.Figueiredo, C. *et al.* Pathogenesis of gastric cancer: genetics and molecular classification. *Molecular Pathogenesis and Signal Trans- duction by Helicobacter pylori* 277–304 (2017).

32.Castro, C., Peleteiro, B. & Lunet, N. Modifiable factors and esophageal cancer: A systematic review of published meta-analyses. *J. Gastroenterol.* **53**, 37–51 (2018).

33.Rajilic-Stojanovic, M. *et al.* Systematic review: gastric microbiota in health and disease. *Aliment. Pharmacol. Therapeutics* **51**, 582–602 (2020).

34.Vinasco, K., Mitchell, H. M., Kaakoush, N. O. & Castaño-Rodríguez, N. Microbial carcinogenesis: Lactic acid bacteria in gastric cancer. *Biochim. Biophys. Acta BBA-Rev. Cancer* **1872**, 188309 (2019).

35.Elliott, D. R. F., Walker, A. W., O'Donovan, M., Parkhill, J. & Fitzgerald, R. C. A non-endoscopic device to sample the oesophageal microbiota: A case-control study. *Lancet Gastroenterol. Hepatol.* **2**, 32–42 (2017).

36.McIlvanna, E., Linden, G. J., Craig, S. G., Lundy, F. T. & James, J. A. Fusobacterium nucleatum and oral cancer: A critical review. *BMC Cancer* **21**, 1–11 (2021).Bronzato, J. D. *et al.* Detection of fusobacterium in oral and head