genes

MDPI

Article

Unraveling the Dysbiosis of Vaginal Microbiome to Understand Cervical Cancer Disease Etiology—An Explainable AI Approach

Karthik Sekaran [1], Rinku Polachirakkal Varghese [1], Mohanraj Gopikrishnan [1], Alsamman M. Alsamman [2], Achraf El Allali [3,*], Hatem Zayed [4] and George Priya Doss C [1,*]

[1] School of Biosciences and Technology, Vellore Institute of Technology, Vellore 632014, India
[2] Molecular Genetics and Genome Mapping Laboratory, Genome Mapping Department, Agricultural Genetic Engineering Research Institute, Cairo 12619, Egypt
[3] African Genome Center, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco
[4] Department of Biomedical Sciences, College of Health Sciences, QU Health, Qatar University, Doha 2713, Qatar
* Correspondence: achraf.elallali@um6p.ma (A.E.A.); georgepriyadoss@vit.ac.in (G.P.D.C.)

Abstract: Microbial Dysbiosis is associated with the etiology and pathogenesis of diseases. The studies on the vaginal microbiome in cervical cancer are essential to discern the cause and effect of the condition. The present study characterizes the microbial pathogenesis involved in developing cervical cancer. Relative species abundance assessment identified *Firmicutes*, *Actinobacteria*, and *Proteobacteria* dominating the phylum level. A significant increase in *Lactobacillus iners* and *Prevotella timonensis* at the species level revealed its pathogenic influence on cervical cancer progression. The diversity, richness, and dominance analysis divulges a substantial decline in cervical cancer compared to control samples. The β diversity index proves the homogeneity in the subgroups' microbial composition. The association between enriched *Lactobacillus iners* at the species level, *Lactobacillus*, *Pseudomonas*, and *Enterococcus* genera with cervical cancer is identified by Linear discriminant analysis Effect Size (LEfSe) prediction. The functional enrichment corroborates the microbial disease association with pathogenic infections such as aerobic vaginitis, bacterial vaginosis, and chlamydia. The dataset is trained and validated with repeated k-fold cross-validation technique using a random forest algorithm to determine the discriminative pattern from the samples. SHapley Additive exPlanations (SHAP), a game theoretic approach, is employed to analyze the results predicted by the model. Interestingly, SHAP identified that the increase in Ralstonia has a higher probability of predicting the sample as cervical cancer. New evidential microbiomes identified in the experiment confirm the presence of pathogenic microbiomes in cervical cancer vaginal samples and their mutuality with microbial imbalance.

Keywords: cervical cancer; eXplainable AI; vaginal microbiome; SHapley Additive exPlanations

Citation: Sekaran, K.; Varghese, R.P.; Gopikrishnan, M.; Alsamman, A.M.; El Allali, A.; Zayed, H.; Doss C, G.P. Unraveling the Dysbiosis of Vaginal Microbiome to Understand Cervical Cancer Disease Etiology—An Explainable AI Approach. *Genes* **2023**, *14*, 936. https://doi.org/10.3390/genes14040936

Academic Editor: Irina Mohorianu

Received: 23 February 2023
Revised: 10 April 2023
Accepted: 12 April 2023
Published: 18 April 2023

1. Introduction

Cancer is a major contributor to mortality and a significant impediment to extending life expectancy. Global predictions indicate that the burden of cancer will increase for at least the next two decades, contributing significantly to the burden of illness [1,2]. Reproductive malignancies constitute a significant cause of female mortality and morbidity worldwide. Cervical cancer is more prevalent in the female reproductive system malignancies, with 569,847 cases per year, ranking it fourth among the malignancies that strike women globally [2,3]. Cervical cancer initially develops in the cervix uteri, and the malignancy transpires slowly overtime.

The key detrimental factor for the preponderance of cervical cancer is exposure to sexually transmitted human papillomavirus (HPV) [4]. If identified at its initial stages,

cervical cancer may be one of the most treatable forms of cancer [5]. The problem is that most patients only seek therapy once the disease has progressed to a late stage. Many potential reasons exist for patients with cervical cancer to seek treatment at a later stage and have a poor prognosis. The paucity of knowledge, cultural issues, the absence of coordinated cancer prevention, as well as inadequate HPV vaccination strategies are a few reasonable factors [6].

HPV infection is a predominant cause of cervical cancer; environmental factors might also significantly impact cancer progression. Epidemiological studies have repeatedly identified smoking as contributing to cervical cancer [7,8]. The microbial communities are one of the elements yet to be substantially researched. The etiology of cervical cancer is multifaceted, and there is less scientific evidence to support the involvement of bacterial groups in cervical carcinogenesis [9,10]. Although microbial diversity is perceived as a sign of health across different body sites, highly diversified vaginal microbiomes are prominently viewed as aberrant or dysbiotic and usually linked to a diseased condition [11,12]. The metagenomic concepts and the transition of high-throughput sequencing analysis have sparked interest in the connection between microbes and various diseases. According to a study by Huang et al., 2014, vaginal microbiome plays a significant role in preserving vaginal homeostasis and limiting the growth of dangerous bacteria [13].

Recent research has evaluated the potential link between cervical cancer and vaginal microbiome [14–19]. Cervical microbiome varies from person to person [20]. It is being investigated as a target for developing novel treatment methods due to mounting evidence that it plays a significant role in the uterine cervix's carcinogenesis process [21,22]. The cervical microbiome is crucial as it possesses the metabolic and enzymatic machinery needed to digest vital vitamins, eliminate harmful substances, fight off infections, support the female genital tract epithelium, and activate and control the immune system [23]. According to earlier research, changes occurred in the cervical microbiota, enhancing the likelihood of carcinogenic development in the cervix. Similar studies demonstrated that altering the cervical microbiome increases the risk of carcinogenic progression [24–26]. Despite the intriguing antecedent results published up to this point, little is still understood about the intricate relationship between cervical dysbiosis and cancer pathogenesis. There is a critical need to compare differences in women with different grades of cervical cancer and their microbial composition to fully understand the microbiome actively involved during cervical cancer pathogenesis. The present study analyzes the vaginal microbial samples of cervical cancer and control groups. Abundance assessment at different taxonomic levels is performed. The α and β diversity are calculated with richness, dominance, and similarity indices of microbial communities between groups. LEfSe analysis detected enriched microbiomes at an LDA score threshold of 3.0. Further, the functional enrichment predicted highly correlated disease association based on the differential microbiomes. SHAP algorithm interpreted the random forest predictions to understand specific microbiomes influencing the results.

2. Materials and Methods

2.1. Data Acquisition

This study intends to compare and analyze the dysbiosis in the vaginal microbiome of cervical cancer patients and healthy individuals. "Cervical cancer" and "Vaginal microbiome" keywords were used to search the NCBI BioProject by applying the filters "Human" as the organism type and "metagenome" as the study type. The vaginal swab samples collected from cervical cancer patients and healthy individuals were sequenced using the 16S rRNA technology to create the final dataset (BioProject ID: PRJNA725946). The vaginal samples were extracted from the genomic DNA using QIAamp DNA Mini Kit and processed with Illumina HiSeq platform at Dalian Medical University, Dalian, China. The samples were labeled according to the patients and the controls. The dataset comprises 65 cervical cancer samples and 54 healthy samples collected using a vaginal swab.

2.2. Bioinformatic Processing and Statistical Analysis

The raw FASTQ files for the vaginal samples (BioProject ID: PRJNA725946) were retrieved from the European Nucleotide Archive (ENA). The single-end reads fetched from the 16S rRNA sequencing method were perused using Quantitative Insights into Microbial Ecology version 2 (QIIME2 v. 2022.8) (https://qiime2.org/ (accessed on 4 December 2022)) [27]. The single-end reads were imported into the QIIME2 and demultiplexed to check the quality of reads. The low-quality reads (Q < 30) were eliminated from the pipeline using trimming and truncation methods. For the single-end reads, the trimming was performed at a beginning position of 0 and abridged at a base length of 240 bp. The DADA2 algorithm was further used to locate and eliminate the chimeric sequences. Following the conventional DADA2 workflow with modifications to accommodate our single-end read data, the 16S sequences were denoised [28].

The sequence's lowest bound of the sampling depth (24,217) was identified to keep all the samples. The sequences with more than 99% similarities were considered Amplicon Sequencing Variants (ASVs). The ASVs considered less than 0.001% of the overall abundance were eliminated to ensure the correctness of the subsequent analysis [29]. The species-level designations were based on precise matching between ASVs and the sequenced reference strain; the taxonomy was determined using the Naïve Bayesian classifier approach using the 16S Silva database (silva-138-99-nb-classifier v. 13_8) [30]. After the aforementioned preprocessing steps, sequences from the phyla of mitochondria and chloroplast were disregarded, as well as those from the kingdoms of Archaea and Eukaryota [31]. The resultant QIIME data, such as the feature and taxonomy tables, were subjected to statistical analysis.

The heterogeneity and uniformity of the microbiota among cervical cancer-affected cases and healthy women were evaluated using α and β diversity analysis [32]. Sequences from each sample were rarefied to a depth of 24,217 to perform the diversity analysis [33]. The samples' α diversity analysis was evaluated using Chao1, Shannon, and Simpson measures based on Wilcoxon rank-sum test [34]. The species differences between the samples were computed using β diversity analysis (PCoA) with Bray Curtis distance metric [35]. The visualization plots for the abovementioned analysis were generated using the micro eco R package [36]. The coalition network was constructed with the igraph R package [37]. Using methods from the igraph package, topographical network characteristics such as centrality and edge weights were also examined.

The differentially represented microbial species between groups at different levels in the taxonomic scale were determined using the LEfSe (http://huttenhower.sph.harvard.edu/galaxy/ (accessed on 5 December 2022)). LDA employs the Kruskal–Wallis approach to determine the traits that show differential abundance among various classes. Using the LEfSe method, variations in microbial abundance between diseased and healthy control groups were determined with a logarithmic LDA score of 4.0. A cladogram and bar graph drawn to show the taxonomic traits are the outputs of the LEfSe model [38]. The functional disease enrichment was performed using the R package MicrobiomeProfiler to study the association between vaginal microbiome and cervical cancer. The microbe–disease enrichment analysis module from the package was utilized to perform the enrichment analysis.

2.3. SHAP Interpretation of Vaginal Microbiome Associated with Cervical Cancer

The collapsed taxonomic table at the species level containing ASVs and taxa information of all the samples was processed with the "DALEX" library in Python [39]. This analysis was intended to show the species identified to have a strong association with cervical cancer alongside complete taxonomic information. SHAP (Shapley Additive Explanations) and DALEX (Descriptive Machine Learning Explanations) are two popular Python libraries used for explainable artificial intelligence (XAI) [40,41]. These libraries provide tools for understanding the behavior of complex machine learning models, such as deep neural networks, decision trees, random forests, and gradient-boosting machines. In this experimental work, the interpretability of the random forest algorithm was evaluated on the vaginal microbial data.

SHAP is a game theoretic approach to explain the output of any machine learning model. It aims to explain the contribution of each input feature to the final model prediction. SHAP computes the Shapley values, which is a measure of the marginal contribution of a feature towards the prediction. Shapley values provide a unified framework for explaining any machine learning model, regardless of its complexity. SHAP also provides visualizations that help understand each feature's importance in the model output. DALEX explains the behavior of machine learning models with the help of visualizations. It provides tools for model-agnostic explanations, feature importance, and model diagnostics.

3. Results

To compare the vaginal microbiome differences between the cervical cancer patients and healthy controls using the ASVs, 119 metagenome sequenced samples were retrieved from the cervical cancer study, including 65 cervical cancer patients (54.6%) and 54 healthy controls (45.3%).

3.1. Characterization of Vaginal Microbiome

After the quality filtering process, there were 5,253,668 reads with a mean value of 44,148 reads per sample. In total, 1973 ASVs were detected after clustering for the sequences at a 99% similarity with the SILVA database. The mean taxon abundance was assessed at different taxonomic levels, such as species, genus, family, class, and phylum, for both cervical cancer and control groups. The top five bacteria belonged to *Firmicutes*, *Actinobacteriota*, *Proteobacteria*, *Bacteroidota*, and *Fusobacteria*, with *Firmicutes* being the most predominant phyla in both groups (Figure 1). The higher taxonomic abundancies at the class level were observed in *Bacilli*, *Actinobacteria*, *Gammaproteobacteria*, *Clostridia*, and *Bacteroidia*, of which *Bacilli* showed greater prevalence (Figure 2). In terms of abundance, Lactobacillus was shown to be the most prevalent, followed by *Gardnerella*, *Streptococcus*, and *Pseudomonas* at the genus level (Figure 3). No cardinal variations were observed in abundance between cervical cancer and healthy control groups at the genus level. *Lactobacillus iners*, *Gardnerella vaginalis*, *Streptococcus agalactiae*, *Streptococcus anginosus*, and *Prevotella timonensis*, among which *Lactobacillus iners* showed higher preponderance in the cervical cancer group at the species level (Figure 4).

Figure 1. Alluvial plot depicting the taxonomic abundance of ASVs associated with the samples at a phylum level.

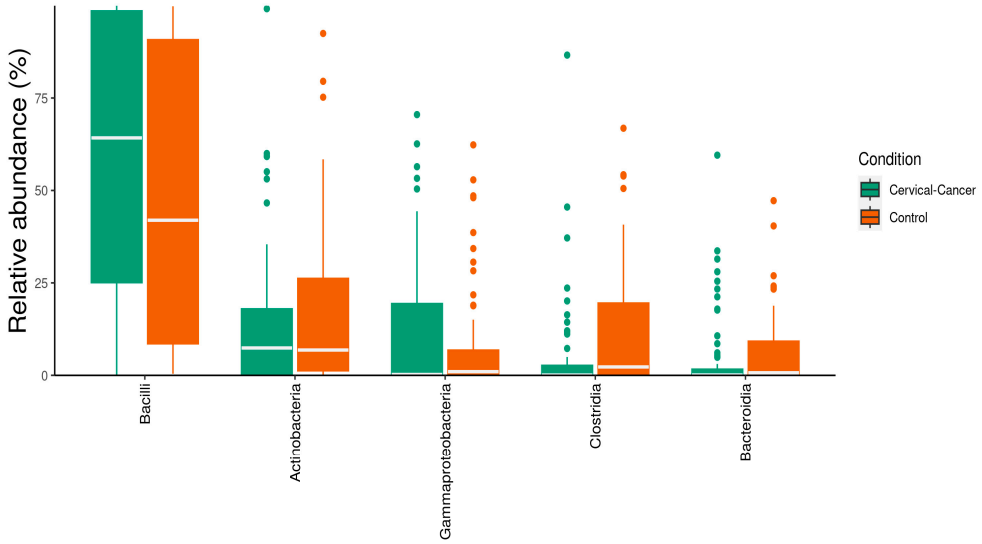

Figure 2. Boxplot illustrating the relatively abundant microbiome of diseased and control groups at the class level. The cervical cancer group is shown in green color, and the healthy control group is shown in orange color.

Figure 3. An illustration of a bar plot showing the relative abundance of diseased and control groups at the genus level (cervical cancer—**left**; healthy control—**right**).

3.2. Dysbiosis of Vaginal Microbiome Associated with Cervical Cancer

Simpson, Shannon, and Chao1 indices were used to understand the complexity of species heterogeneity between the two groups. The species richness within the samples can be reflected using Chao1, whereas Shannon and Simpson indices depict the species diversity within a community (species richness and diversity). The Chao1 measure is considerably higher for healthy control than for the cervical cancer group. As per the findings, species richness is substantially higher in healthy controls. The Shannon and Simpson measures show higher indices for the healthy control group than the cervical cancer group (Figure 5).

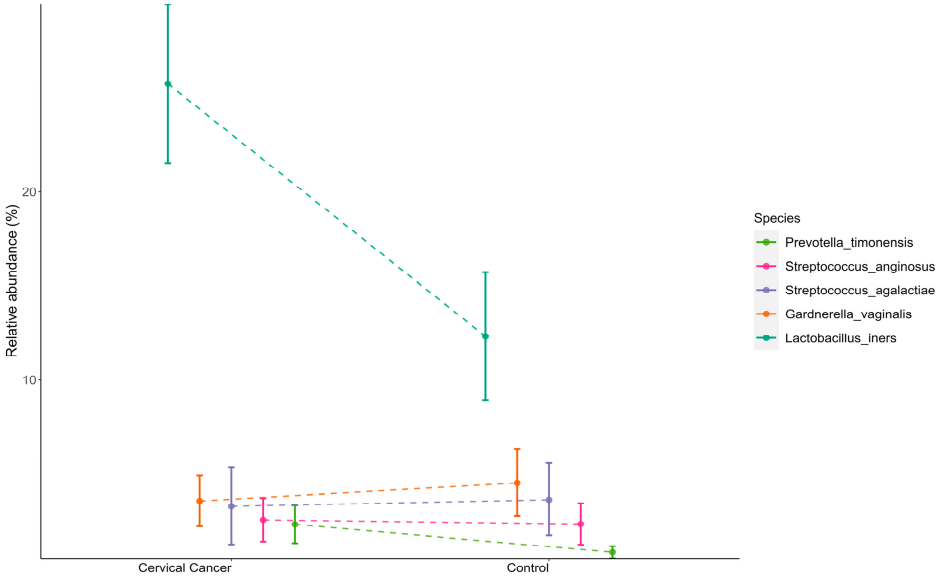

Figure 4. Line plot representing the taxonomic abundance at species level across the different groups (cervical cancer—**left**; healthy control—**right**).

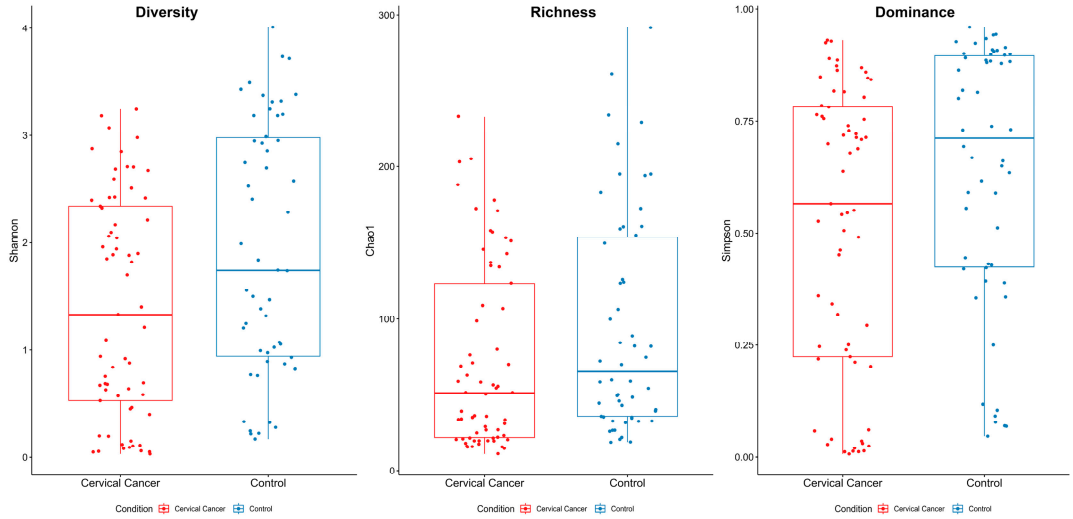

Figure 5. α diversity indexes are plotted as boxplots. α diversity indices are composite indices that capture consistency and abundance. The Shannon and Simpson indices reflect ASV diversity in samples, and the Chao1 measure reflects the ASV abundance in samples.

The vaginal microbiota diversity among the two groups was compared using the Bray–Curtis distance measure. The microbial makeup of each group can be represented using a Principal coordinate analysis (PCoA) plot (Figure 6). In PCOA plots, the samples closer to each other resemble similar microbial communities. In the PCoA plot, the two coordinates (PCo1 and PCo2) account for 34.7% of the variation.

The coalition network can be used to depict the associativity between microorganisms present within a group or a community. The PCoA plot indicates a significant distinction among the vaginal microbial communities of cervical cancer and healthy control groups (p-value: 0.001, R2: 0.027, F-value: 3.269). The igraph bipartite approach was used to identify the connections among different microbes at the class level. Alphaproteobacteria were identified as the key taxon within the network that formed pairwise co-occurrence networks with the other microbes, particularly with *Gammaproteobacteria*, *Bacteroidia*, *Actinobacteria*, and *Bacilli* (Figure 7).

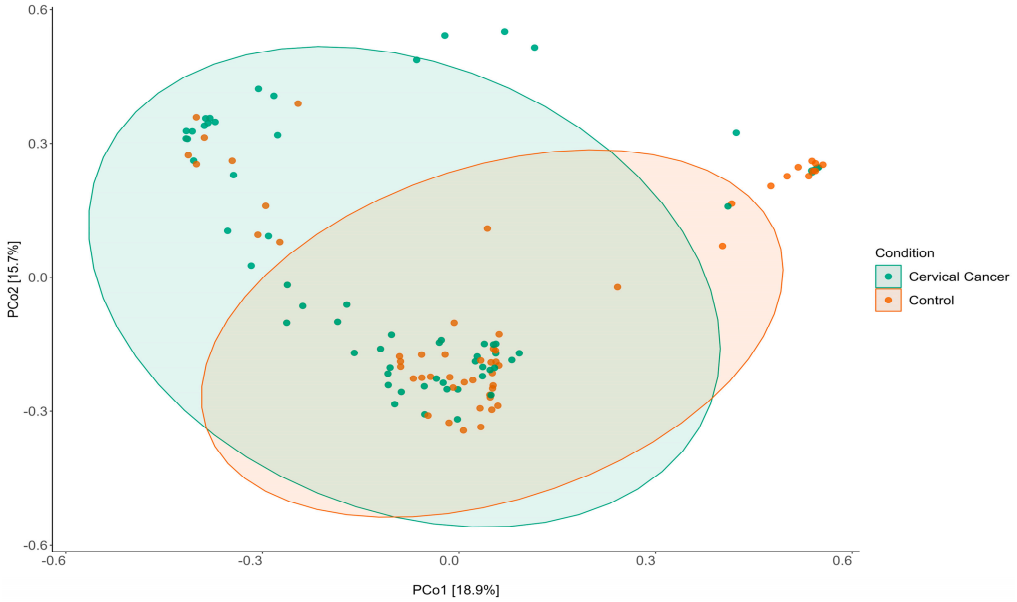

Figure 6. PCOA plots of β diversity of vaginal microbiota based on Bray−Curtis distance measure. The ellipses represent the two groups. The cervical cancer is shown in green color, whereas the healthy control group is shown in orange color.

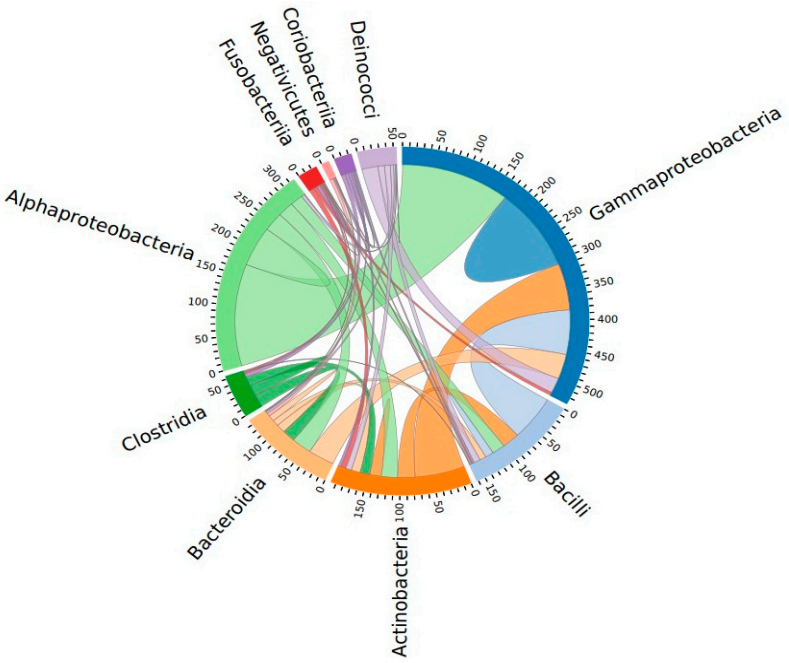

Figure 7. The chord diagram displays the network of 10 candidates that co-occur in a pairwise sequence. Each sector of the circle represents a node (i.e., taxon) in the network, and its width reflects the sum of the co-occurrences between each taxon.

LEfSe assessment identifies the microbial abundance of cervical cancer patients and healthy control group from the vaginal microbiome. The LEfSe profiling shows variations between cervical cancer and healthy control groups at various taxon levels with a threshold LDA core of 4.0 (Figure 8). In cervical cancer patients, the cladogram shows a significant abundance of *Lactobacillus iners*, *Pseudomonadaceae*, *Enterococacceae*, and *Entomoplasmatales*, whereas *Proteobacteria*, *Actinobacteria*, and *Bacteroidota* are displayed in the healthy control (Figure 8).

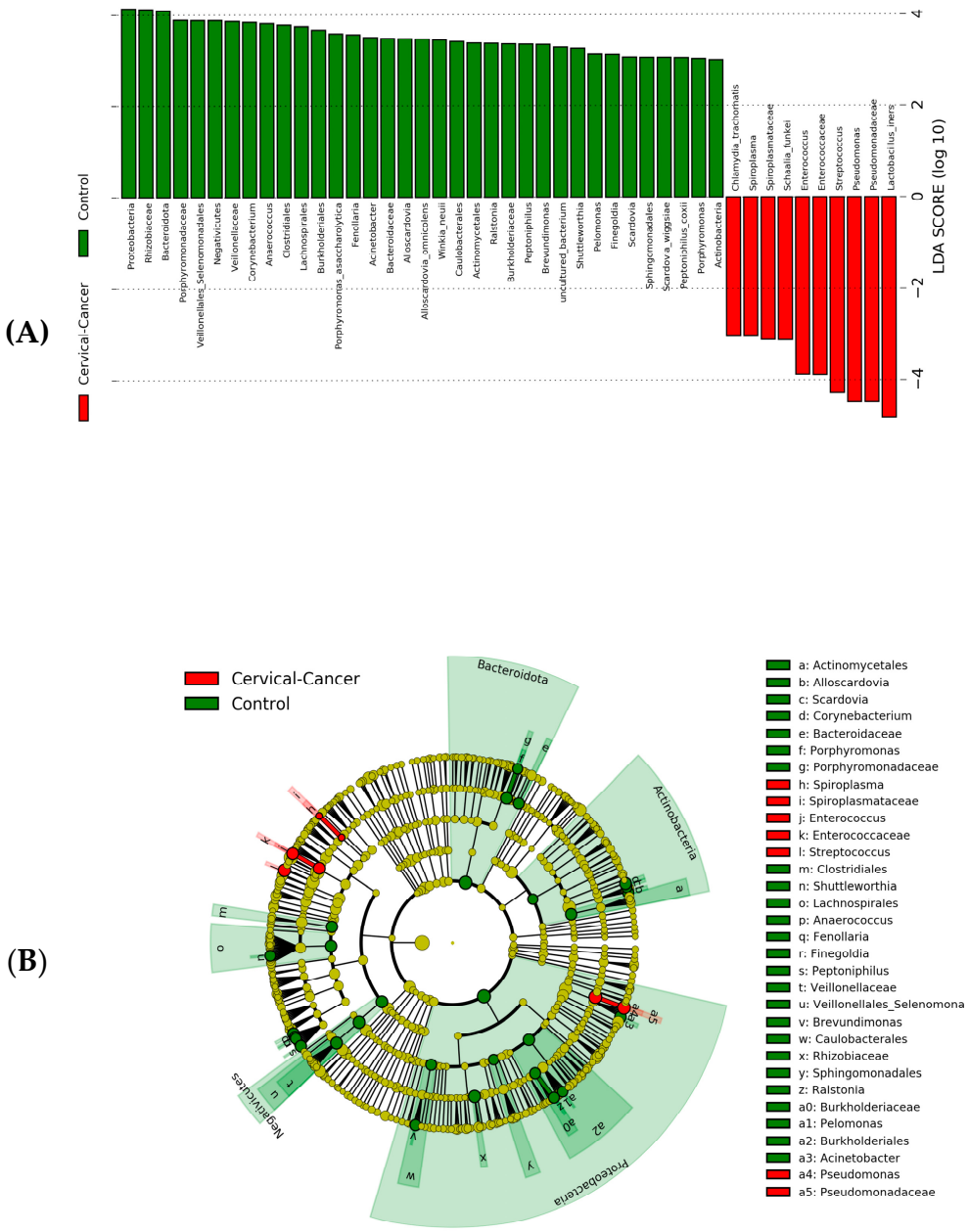

Figure 8. Linear discriminant analysis (LDA) effect size (LEfSe) analysis for vaginal microbiota abundance in cervical cancer and healthy control groups. (**A**) The Bar plot from LEfSe analysis indicates the enriched bacteria that are associated with cervical cancer (red) and healthy control (green) groups. (**B**) A phylogenetic cladogram plot from LEfSE analysis representing the differentially abundant taxa at various taxonomic levels (LDA score > 3.0).

The differential expressed taxa were detected using MicrobiomeProfiler to identify the bacterial strains enriched in the vaginal microbiota of cervical cancer patients. The disbiome database was selected for microbiome disease enrichment analysis, for which the taxon IDs of identified bacterial strains (135) were provided as input (Table S1). The microbial strains were determined to be associated with eight diseases, of which the microbial enrichment were highly associated with Aerobic vaginitis, Bacterial vaginosis, and Chlamydia, respectively. The functional enrichment outcome between cervical cancer and healthy vaginal microbiome samples is represented in Figure 9.

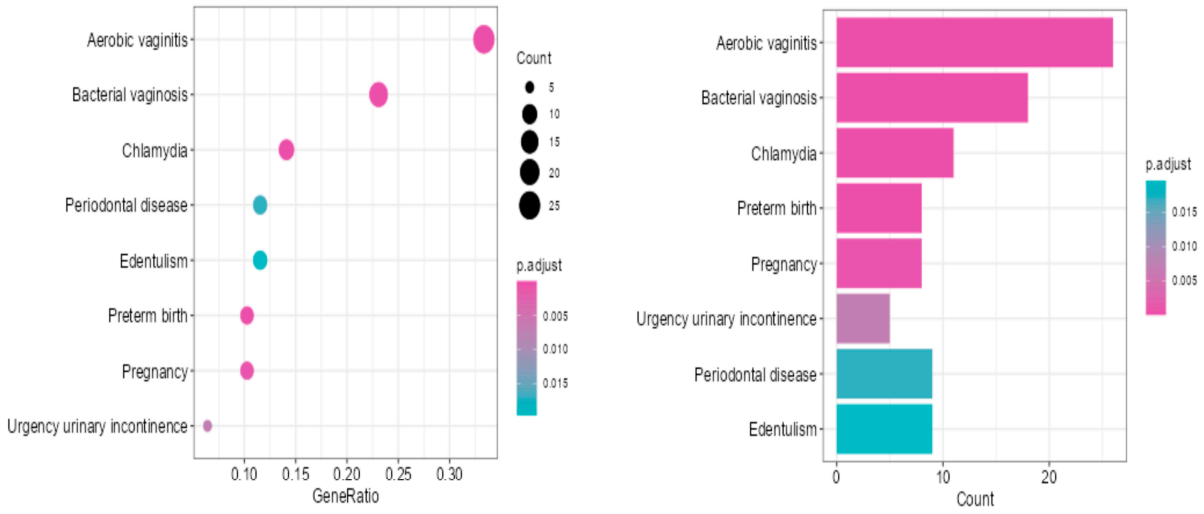

Figure 9. Comparative disease–microbiome enrichment analysis of vaginal microbiota depicted as line plot and bar plot. A total of 77 significantly different bacterial taxa were reported in the enrichment analysis.

3.3. Explaining the Model Predictions through SHAP

Interpreting "black-box" mathematical models is pivotal to understanding complex biological outcomes. Traditional machine learning algorithms generate results based on intuitive, logical assessments derived through mathematical models. However, the reason for every model prediction is unknown due to the higher level of abstraction and deeper computing process. It is also arduous to analyze each step of interminable calculation performed by the algorithms. Explainable Artificial Intelligence (XAI), a sophisticated algorithmic approach, was developed by the Defense of Advanced Projects Research Agency (DARPA). It is intended to develop self-explainable human understandable models while maintaining higher-level performance. Shapley Additive Explanations (SHAP), a game theoric approach-based framework, conduct interpretable predictions from the results of any trained machine learning model. This method assigns importance to a particular sample prediction variable based on the Shapley values. The average marginal contribution of every feature score over all other possible coalitions calculates it. DALEX provides tools for creating various model-agnostic explanations, such as feature importance plots, partial dependence plots, and accumulated local effects plots. The SHAP value plot, breakdown, and ROC curve results are visualized using DALEX.

The microbiome dataset contains a taxonomic hierarchy from Kingdom to Species-level of each column as a feature vector with 594 taxons in total, and 119 rows represent individual samples. Random forest, an ensemble-based bagging model, is trained with the data to numerically understand the discriminative pattern between microbiomes of cervical cancer and control samples. The model performance is evaluated through k-fold cross-validation (K = 10) and repeated k-fold cross-validation with five repeats. The k-fold and repeated k-fold CV scores are 0.926 and 0.971, respectively, and share no big difference between the results (Supplementary file). The resultant model of repeated k-fold CV is inputted into the SHAP model to understand the predictions. Two samples from the dataset of each study group are randomly drawn for interpretation. The SHAP results of cervical cancer and control samples are depicted in Figures 10 and 11, respectively. The *X*-axis represents the taxonomic label, and the contribution of each feature is provided as a probability score in the *Y*-axis. The top bar plot of Figures 10 and 11 visualizes the importance of each feature contributing to predicting a particular class in terms of SHAP values.

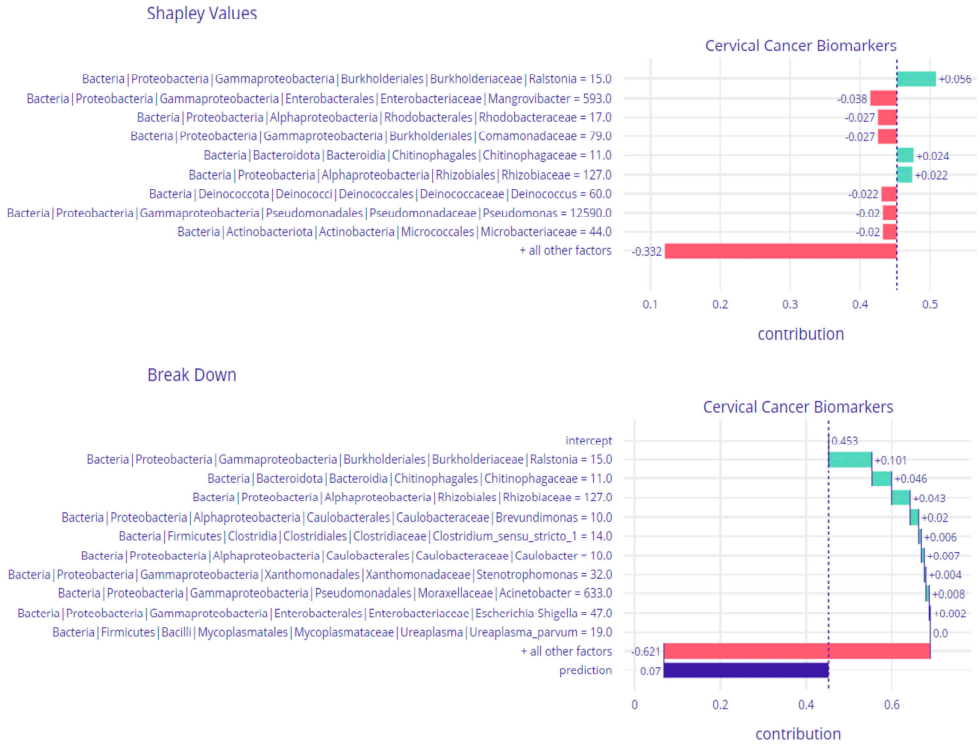

Figure 10. SHAP explanations for the cervical cancer group, the top bar plot depicts the importance of each feature in terms of SHAP values. The bottom plot represents the feature breakdown. The red bar signifies a declining pattern, whereas the green bar shows an increase in the average response for each feature.

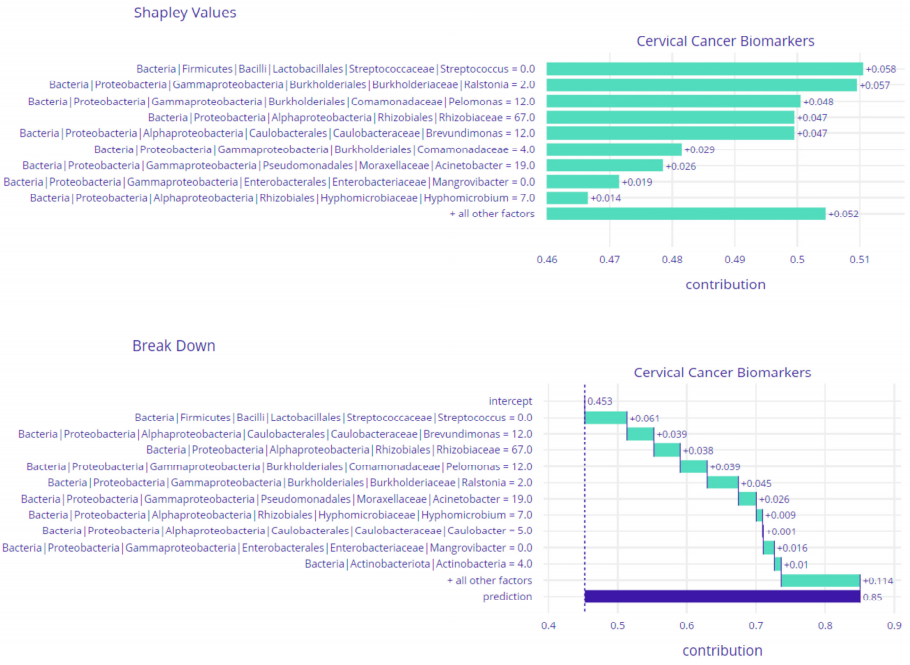

Figure 11. The SHAP explanations for the healthy control group, where the bar plots represent feature significance using SHAP values and feature breakdowns, respectively. The green patterns illustrate the substantial increase in average response for each feature.

Similarly, the bottom bar plot provides each feature breakdown contributing to the correct prediction of the corresponding sample class. Each feature's negative and positive impact on the predictions is represented in red and green. The green bar indicates the

increase in the average response of each feature, whereas the red bar denotes the decreasing pattern. The intercept value is the average response score; in the current model, it is 0.453.

The increased *Ralstonia* at the genus level, *Chitinophagaceae*, and *Rhizobiaceae* Family level positively impacted the sample prediction as cervical cancer, provided at the top of Figure 10. The breakdown figure at the bottom provides the positive contribution of each microbiome in the prediction. This inference exhibits the importance of the microbiomes mentioned above in classifying cervical cancer individuals. The analysis of the control sample in Figure 11 determined that the decreased count of *Streptococcus*, *Ralstonia*, *Pseudomonas*, and *Brevundimonas* at the genus level positively correlated with the control sample.

Ralstonia and *Rhizobiaceae* were observed in both predictions. However, the decrease in the count of these microbiomes contributed to the control sample prediction. The prediction probability confidence of the model on the cervical cancer sample is 0.07, and the control sample is 0.85, with class label values 0 and 1, respectively. Figure 12 depicts the ROC curve of the random forest model at the top, with a score of 1. The reverse cumulative distribution curve at the bottom indicates that most residuals fall below 0.1. This phenomenon occurs when the dataset contains many features, assigning varying contributions to every feature.

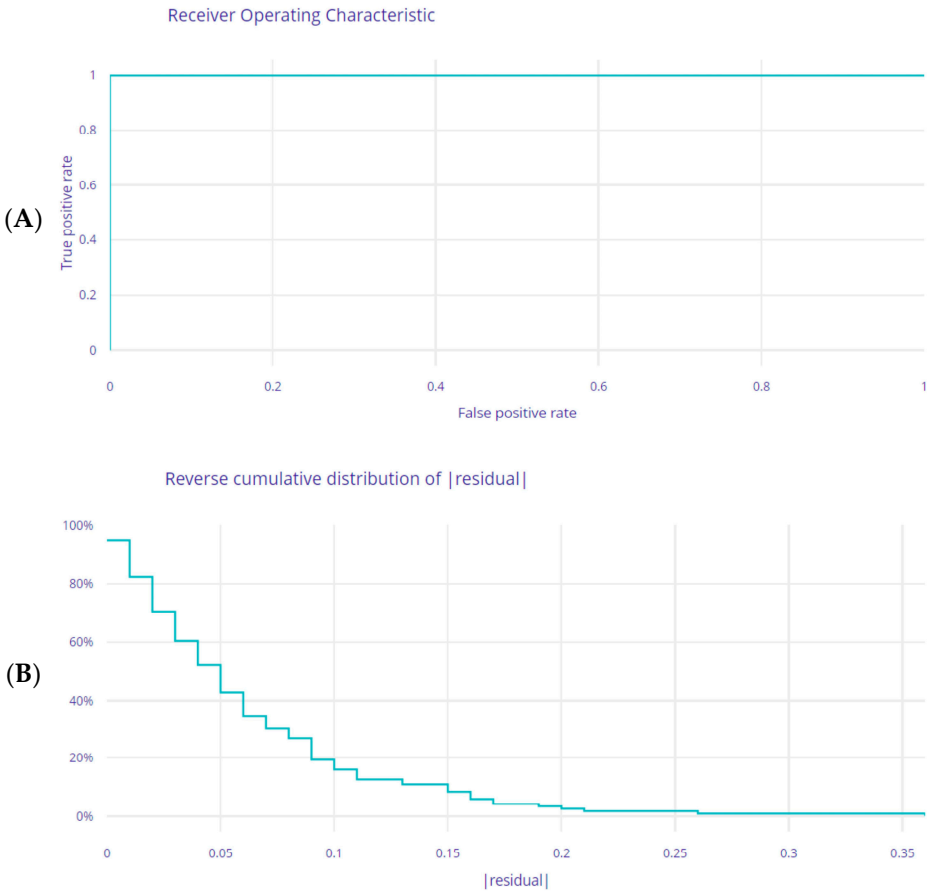

Figure 12. Model construction and feature screening using machine learning algorithms. (**A**) ROC curve of random forest model construction with a residual value of 1. (**B**) The reverse cumulative distribution curve plot to show the residual distribution of the random forest model.

4. Discussion

Characterization of the microbiome is essential to untangle the disease etiology. Microbial dysbiosis is a crucial factor associated with disease dynamics, also evident in accurate diagnosis of the condition. This study analyzed the vaginal microbiome of 65 cervical cancer and 54 healthy samples to discern microbial pathogenicity. The taxon abundance assessment at different levels determined unique microbial patterns exhibiting clear discrimination between the case and control groups. *Firmicutes*, *Actinobacteria*, and

Proteobacteria, are abundant at the Phylum level. *Lactobacillus* genera are elevated when compared to *Gardnerella* and *Streptococcus*. In much literature, the influence of *Lactobacillus* on cervical cancer is reported [42,43]. *Lactobacillus iners* showed higher abundance in cervical samples over control (Figure 4). The oncogenic nature of *Lactobacillus iners* in cervical cancer was delineated in a microbial study [44]. Other abundant species, such as *Prevotella timonensis* [45,46], *Gardnerella vaginalis* [47], and *Streptococcus anginosus* [48], confirmed microbial pathogenicity.

The diversity and richness analysis identified a decline in the cervical cancer microbial community, calculated by Shannon and Chao index. The Bray–Curtis distance measure was used to quantify the compositional dissimilarity of the microbiome, visualized using PCoA. The plot displayed a distinct cluster pattern among the vaginal microbial communities of cervical cancer and healthy control groups with *p*-value: 0.001, R2: 0.027, and F-value: 3.269 (Figure 6). LEfSe predicted enriched taxonomical units at a different level. *Lactobacillus iners* ranked top, followed by *Pseudomonas*, *Streptococcus*, and *Enterococcus*, describing the pathogenic association with cervical cancer. *Proteobacteria*, *Rhizobiaceae*, and *Bacteriodota* were highly enriched in the control group.

The differentially expressed taxa were calculated to perform disease-functional enrichment of microbiomes. The disease association of the enlisted taxa reported aerobic vaginitis, bacterial vaginosis, and chlamydia. Prolonged exposure to the pathogenic bacterial environment increases the risk of developing cervical cancer [49]. Another dimension of this study scrutinized the influence of each microbe contributing to the discrimination of cervical cancer and control samples. It examined the importance of each feature and its impact on prediction through SHAP values. The dataset was trained with a random forest ensemble classification algorithm. The prediction result of the model was interpreted using the SHAP algorithm. Increased *Ralstonia* impacted the prediction of the sample as cervical cancer with a higher probability (0.056) [50].

Conversely, the highly pathogenic taxa, *Streptococcus* [51], has a minor abundance contributing to the prediction (0.058) of the control sample, followed by *Ralstonia* (0.057). The reverse cumulative distribution curve indicates that the features lie below 0.1, impacting the predictions (Figure 12). The lesser value is due to many features (594) in the database. This study unveiled many potential pathogenic vaginal microbiomes causing a detrimental effect on individuals. Meanwhile, there exist many factors involved in the disease condition. Multi-omic studies on cervical cancer will further broaden the understanding of the disease etiology. Clinical informatics, combined with artificial intelligence, makes personalized medicine possible in the near future to treat complex diseases through effective mechanisms.

5. Conclusions

This study identified the dominance of *Lactobacillus iners* species in the vaginal microbiome of cervical cancer samples. The imbalance in microbial distribution is observed during α diversity analysis. *Lactobacillus*, *Gardnerella*, *Pseudomonas,* and *Enterococcus* are abundant at the genus level in cervical cancer. The microbiome disease association enrichment detects increased susceptibility with aerobic vaginitis, bacterial vaginosis, and chlamydia. These diseases have a direct coalition with cervical cancer and other severe vaginal infections. The discriminative evidence to classify healthy and cervical cancer group samples is deliberated with the SHAP model. The explainable approach identifies *Ralstonia* as a microbial predictor marker. The increased composition of *Ralstonia* impels the model to predict the sample as cervical cancer. Though *Ralstonia* is not reported as highly prevalent in cervical cancer, this inference unveils the decisive characteristics of the marker. Thus, the current findings invigorate the development of probiotics as targeted therapeutics for effective treatment. The following limitation is identified and reported in the present work. This study delineates the microbiome information of a single dataset, and though it is valid, a comparative analysis cannot be conducted. In the future, this study could be further extended by adding more datasets to demonstrate and benchmark the results, thereby ensuring in-depth validation of the findings.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14040936/s1, Table S1: Functional Enrichment Analysis, Supplementary file: Repeated k-fold cross-validation results.

Author Contributions: K.S., R.P.V., M.G., A.E.A., A.M.A., H.Z. and G.P.D.C. were involved in the study's design. The data collection and experiment involved K.S., R.P.V. and M.G. K.S., R.P.V., A.M.A. and M.G. acquired, analyzed and interpreted the results. A.E.A. and G.P.D.C. supervised the entire study. K.S., R.P.V., A.M.A., A.E.A. and M.G. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available with the corresponding author AEA and GPDC. The code is available in the following github link: https://github.com/karthiksekaran/microbiome-AI, accessed on 4 December 2022.

Acknowledgments: The authors would like to thank the authorities of the Vellore Institute of Technology, India for providing the necessary support in completing the manuscript. The authors acknowledge the Indian Council of Medical Research (ICMR), the Government of India agency, for the research grants No. BMI/12(13)/2021, ID No: 2021-6359 and No. VIR/COVID-19/31/2021/ECD-I, ID. NO: 2021-5570. The authors acknowledge the African Supercomputing Center at Mohamed VI Polytechnic University for the supercomputing resources (https://ascc.um6p.ma/, accessed on 4 December 2022) made available for conducting the research reported in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Foreman, K.J.; Marquez, N.; Dolgert, A.; Fukutaki, K.; Fullman, N.; McGaughey, M.; Pletcher, M.A.; Smith, A.E.; Tang, K.; Yuan, C.-W.; et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and territories. *Lancet* **2018**, *392*, 2052–2090. [CrossRef] [PubMed]
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
3. Pimple, S.; Mishra, G. Cancer cervix: Epidemiology and disease burden. *CytoJournal* **2022**, *19*, 21. [CrossRef]
4. William, W.; Ware, A.; Basaza-Ejiri, A.H.; Obungoloch, J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Comput. Methods Programs Biomed.* **2018**, *164*, 15–22. [CrossRef] [PubMed]
5. Devi, B.C.R.; Tang, T.S.; Corbex, M. Reducing by half the percentage of late-stage presentation for breast and cervix cancer over 4 years: A pilot study of clinical downstaging in Sarawak, Malaysia. *Ann. Oncol.* **2007**, *18*, 1172–1176. [CrossRef] [PubMed]
6. Hicks, M.L.; Yap, O.W.S.; Matthews, R.; Parham, G. Disparities in cervical cancer screening, treatment and outcomes. *Ethn. Dis.* **2006**, *16* (Suppl. S3), S3-63–S3-66.
7. Plummer, M.; Herrero, R.; Franceschi, S.; Meijer, C.J.L.M.; Snijders, P.; Bosch, F.X.; de Sanjosé, S.; Muñoz, N. Smoking and cervical cancer: Pooled analysis of the IARC multi-centric case–control study. *Cancer Causes Control* **2003**, *14*, 805–814. [CrossRef]
8. Roura, E.; Castellsagué, X.; Pawlita, M.; Travier, N.; Waterboer, T.; Margall, N.; Bosch, F.X.; de Sanjosé, S.; Dillner, J.; Gram, I.T.; et al. Smoking as a major risk factor for cervical cancer and pre-cancer: Results from the EPIC cohort. *Int. J. Cancer* **2014**, *135*, 453–466. [CrossRef]
9. Adebamowo, S.N.; Dareng, E.O.; Famooto, A.O.; Offiong, R.; Olaniyan, O.; Obende, K.; Adebayo, A.; Ologun, S.; Alabi, B.; Achara, P.; et al. Cohort Profile: African Collaborative Center for Microbiome and Genomics Research's (ACCME's) Human Papillomavirus (HPV) and Cervical Cancer Study. *Int. J. Epidemiol.* **2017**, *46*, 1745–1745j. [CrossRef]
10. Seo, S.-S.; Oh, H.Y.; Lee, J.-K.; Kong, J.-S.; Lee, D.O.; Kim, M.K. Combined effect of diet and cervical microbiome on the risk of cervical intraepithelial neoplasia. *Clin. Nutr.* **2016**, *35*, 1434–1441. [CrossRef]
11. Amabebe, E.; Anumba, D.O.C. The Vaginal Microenvironment: The Physiologic Role of Lactobacilli. *Front. Med.* **2018**, *5*, 181. [CrossRef]
12. Mitra, A.; MacIntyre, D.A.; Marchesi, J.R.; Lee, Y.S.; Bennett, P.R.; Kyrgiou, M. The vaginal microbiota, human papillomavirus infection and cervical intraepithelial neoplasia: What do we know and where are we going next? *Microbiome* **2016**, *4*, 58. [CrossRef] [PubMed]
13. Huang, B.; Fettweis, J.M.; Brooks, J.P.; Jefferson, K.K.; Buck, G.A. The changing landscape of the vaginal microbiome. *Clin. Lab. Med.* **2014**, *34*, 747–761. [CrossRef] [PubMed]

14. Audirac-Chalifour, A.; Torres-Poveda, K.; Bahena-Román, M.; Téllez-Sosa, J.; Martínez-Barnetche, J.; Cortina-Ceballos, B.; López-Estrada, G.; Delgado-Romero, K.; Burguete-García, A.I.; Cantú, D.; et al. Cervical Microbiome and Cytokine Profile at Various Stages of Cervical Cancer: A Pilot Study. *PLoS ONE* **2016**, *11*, e0153274. [CrossRef]
15. Chase, D.; Goulder, A.; Zenhausern, F.; Monk, B.; Herbst-Kralovetz, M. The vaginal and gastrointestinal microbiomes in gynecologic cancers: A review of applications in etiology, symptoms and treatment. *Gynecol. Oncol.* **2015**, *138*, 190–200. [CrossRef]
16. Łaniewski, P.; Barnes, D.; Goulder, A.; Cui, H.; Roe, D.J.; Chase, D.M.; Herbst-Kralovetz, M.M. Linking cervicovaginal immune signatures, HPV and microbiota composition in cervical carcinogenesis in non-Hispanic and Hispanic women. *Sci. Rep.* **2018**, *8*, 7593. [CrossRef]
17. Łaniewski, P.; Ilhan, Z.E.; Herbst-Kralovetz, M.M. The microbiome and gynaecological cancer development, prevention and therapy. *Nat. Rev. Urol.* **2020**, *17*, 232–250. [CrossRef]
18. Mitra, A.; MacIntyre, D.A.; Lee, Y.S.; Smith, A.; Marchesi, J.R.; Lehne, B.; Bhatia, R.; Lyons, D.; Paraskevaidis, E.; Li, J.V.; et al. Cervical intraepithelial neoplasia disease progression is associated with increased vaginal microbiome diversity. *Sci. Rep.* **2015**, *5*, 16865. [CrossRef]
19. Cheng, L.; Norenhag, J.; Hu, Y.O.O.; Brusselaers, N.; Fransson, E.; Ährlund-Richter, A.; Guðnadóttir, U.; Angelidou, P.; Zha, Y.; Hamsten, M.; et al. Vaginal microbiota and human papillomavirus infection among young Swedish women. *Npj Biofilm. Microbiomes* **2020**, *6*, 39. [CrossRef]
20. Klein, C.; Kahesa, C.; Mwaiselage, J.; West, J.T.; Wood, C.; Angeletti, P.C. How the Cervical Microbiota Contributes to Cervical Cancer Risk in Sub-Saharan Africa. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 23. [CrossRef]
21. Brusselaers, N.; Shrestha, S.; Wijgert, J.v.d.; Verstraelen, H. Vaginal dysbiosis and the risk of human papillomavirus and cervical cancer: Systematic review and meta-analysis. *Am. J. Obstet. Gynecol.* **2019**, *221*, 9–18.e8. [CrossRef] [PubMed]
22. Ravilla, R.; Coleman, H.; Chan, L.; Chow, C.-E.; Fuhrman, B.; Greenfield, W.; Robeson, M.S.; Iverson, K.; Spencer, H.J.; Nakagawa, M. Cervical microbiome role in outcomes of therapeutic HPV vaccination for cervical intraepithelial neoplasia. *J. Clin. Oncol.* **2018**, *36* (Suppl. S15), 3099. [CrossRef]
23. Tango, C.N.; Seo, S.-S.; Kwon, M.; Lee, D.-O.; Chang, H.K.; Kim, M.K. Taxonomic and Functional Differences in Cervical Microbiome Associated with Cervical Cancer Development. *Sci. Rep.* **2020**, *10*, 9720. [CrossRef]
24. Arokiyaraj, S.; Seo, S.S.; Kwon, M.; Lee, J.K.; Kim, M.K. Association of cervical microbial community with persistence, clearance, and negativity of Human Papillomavirus in Korean women: A longitudinal study. *Sci. Rep.* **2018**, *8*, 15479. [CrossRef]
25. Khan, I.; Nam, M.; Kwon, M.; Seo, S.; Jung, S.; Han, J.S.; Hwang, G.-S.; Kim, M.K. LC/MS-Based Polar Metabolite Profiling Identified Unique Biomarker Signatures for Cervical Cancer and Cervical Intraepithelial Neoplasia Using Global and Targeted Metabolomics. *Cancers* **2019**, *11*, 511. [CrossRef] [PubMed]
26. Kwon, M.; Seo, S.-S.; Kim, M.K.; Lee, D.O.; Lim, M.C. Compositional and Functional Differences between Microbiota and Cervical Carcinogenesis as Identified by Shotgun Metagenomic Sequencing. *Cancers* **2019**, *11*, 309. [CrossRef] [PubMed]
27. Caporaso, J.G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.D.; Costello, E.K.; Fierer, N.; Peña, A.G.; Goodrich, J.K.; Gordon, J.I.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7*, 335–336. [CrossRef]
28. Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]
29. Yang, X.; He, L.; Yan, S.; Chen, X.; Que, G. The impact of caries status on supragingival plaque and salivary microbiome in children with mixed dentition: A cross-sectional survey. *BMC Oral Health* **2021**, *21*, 319. [CrossRef]
30. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef]
31. Kozich, J.J.; Westcott, S.L.; Baxter, N.T.; Highlander, S.K.; Schloss, P.D. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl. Environ. Microbiol.* **2013**, *79*, 5112–5120. [CrossRef] [PubMed]
32. Willis, A.D. Rarefaction, Alpha Diversity, and Statistics. *Front. Microbiol.* **2019**, *10*, 2407. [CrossRef] [PubMed]
33. Cameron, E.S.; Schmidt, P.J.; Tremblay, B.J.-M.; Emelko, M.B.; Müller, K.M. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Sci. Rep.* **2021**, *11*, 22302. [CrossRef] [PubMed]
34. Thukral, A.K. A review on measurement of Alpha diversity in biology. *Agric. Res. J.* **2017**, *54*, 1. [CrossRef]
35. Lozupone, C.A.; Hamady, M.; Kelley, S.T.; Knight, R. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Appl. Environ. Microbiol.* **2007**, *73*, 1576–1585. [CrossRef]
36. Liu, C.; Cui, Y.; Li, X.; Yao, M. Microeco: An R package for data mining in microbial community ecology. *FEMS Microbiol. Ecol.* **2021**, *97*, fiaa255. [CrossRef]
37. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **2006**, *1695*, 1–9.
38. Segata, N.; Izard, J.; Waldron, L.; Gevers, D.; Miropolsky, L.; Garrett, W.S.; Huttenhower, C. Metagenomic biomarker discovery and explanation. *Genome Biol.* **2011**, *12*, R60. [CrossRef]
39. Baniecki, H.; Kretowicz, W.; Piatyszek, P.; Wisniewski, J.; Biecek, P. Dalex: Responsible machine learning with interactive explainability and fairness in Python. *J. Mach. Learn. Res.* **2021**, *22*, 9759–9765.

40. Fidel, G.; Bitton, R.; Shabtai, A. When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
41. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
42. Yang, X.; Da, M.; Zhang, W.; Qi, Q.; Zhang, C.; Han, S. Role of Lactobacillus in cervical cancer. *Cancer Manag. Res.* **2018**, *10*, 1219. [CrossRef] [PubMed]
43. Kyrgiou, M.; Moscicki, A.B. Vaginal microbiome and cervical cancer. In *Seminars in Cancer Biology*; Academic Press: Cambridge, MA, USA, 2022.
44. Colbert, L.E.; Karpinets, T.V.; El Alam, M.B.; Lynn, E.J.; Sammouri, J.; Lo, D.; Elnaggar, J.H.; Wang, R.; Harris, T.A.; Yoshida-Court, K.; et al. Cancer-associated *Lactobacillus iners* are genetically distinct and associated with chemoradiation resistance in cervical cancer. *medRxiv* **2022**. [CrossRef]
45. Chambers, L.M.; Bussies, P.; Vargas, R.; Esakov, E.; Tewari, S.; Reizes, O.; Michener, C. The microbiome and gynecologic cancer: Current evidence and future opportunities. *Curr. Oncol. Rep.* **2021**, *23*, 92. [CrossRef] [PubMed]
46. Jain, A.; Shrivastava, S.K.; Joy, L. Cervicovaginal microbiota and HPV-induced cervical cancer. In *Immunopathology, Diagnosis and Treatment of HPV Induced Malignancies*; Academic Press: Cambridge, MA, USA, 2022; pp. 81–97.
47. Raffone, A.; Travaglino, A.; Angelino, A.; Esposito, R.; Orlandi, G.; Toscano, P.; Mollo, A.; Insabato, L.; Sansone, M.; Zullo, F. *Gardnerella vaginalis* and *Trichomonas vaginalis* infections as risk factors for persistence and progression of low-grade precancerous cervical lesions in HIV-1 positive women. *Pathol.-Res. Pract.* **2021**, *219*, 153349. [CrossRef]
48. Liu, H.; Liang, H.; Li, D.; Wang, M.; Li, Y. Association of Cervical Dysbacteriosis, HPV Oncogene Expression, and Cervical Lesion Progression. *Microbiol. Spectr.* **2022**, *10*, e00151-22. [CrossRef]
49. Wei, W.; Xie, L.Z.; Xia, Q.; Fu, Y.; Liu, F.Y.; Ding, D.N.; Han, F.J. The role of vaginal microecology in the cervical cancer. *J. Obstet. Gynaecol. Res.* **2022**, *48*, 2237–2254. [CrossRef]
50. Lin, S.; Zhang, B.; Lin, Y.; Lin, Y.; Zuo, X. Dysbiosis of Cervical and Vaginal Microbiota Associated with Cervical Intraepithelial Neoplasia. *Front. Cell. Infect. Microbiol.* **2022**, *12*, 767693. [CrossRef]
51. Wang, Z.; Xiao, R.; Huang, J.; Qin, X.; Hu, D.; Guo, E.; Liu, C.; Lu, F.; You, L.; Sun, C.; et al. The diversity of vaginal microbiota predicts neoadjuvant chemotherapy responsiveness in locally advanced cervical cancer. *Microb. Ecol.* **2022**, *84*, 302–313. [CrossRef]