Machine Learning Identification of Gut Microbiome Signatures Associated with Thyroid Cancer

# 1 Abstract

To explore potential gut microbiota markers associated with thyroid cancer using bioinformatics and machine learning techniques.

Methods: We analyzed gut microbiome data from the NCBI project SRP151288, utilizing the Kraken2 tool for sequence classification and generating Operational Taxonomic Unit (OTU) tables. Multiple machine learning models, including generalized linear models, distributed random forests, extremely randomized trees, and deep learning, were employed on the H2O platform to identify significant microbial features. Non-parametric Wilcoxon tests were conducted to validate these features.

Results: Several potential microbial markers for thyroid cancer were identified, including OTU965\_g\_\_Moraxella, OTU743\_g\_\_Sutterella, OTU2419\_g\_\_Emergencia, OTU2418\_g\_\_Aminipila, and OTU2413\_g\_\_Christensenella. Sutterella showed significantly higher abundance in the healthy control group, while Emergencia, Lactococcus, and Carnobacterium exhibited enrichment trends in thyroid cancer patients.

Conclusion: This study provides new insights into the relationship between gut microbiota and thyroid cancer, identifying potential biomarkers for diagnosis and treatment. These findings contribute to our understanding of the gut-thyroid axis and may guide future research in thyroid cancer pathogenesis and personalized medicine approaches.

# 2 Introduction

In modern medical research, the role of the human microbiome has garnered widespread attention, particularly the impact of gut microbiota on host health[1]. Extensive studies have demonstrated that the gut microbiome is intricately linked to the host’s metabolism, immune function, and the development of various diseases, including cancer. As the most common endocrine malignancy, thyroid cancer has shown a continuous increase in incidence in recent years, prompting researchers to explore novel biomarkers for improved diagnosis, prognosis assessment, and treatment strategies[2].

This study aims to explore potential markers of gut microbiota in thyroid cancer. Through in-depth analysis of gut microbiome data associated with thyroid cancer, we attempt to uncover the potential relationship between intestinal microbes and thyroid cancer. We obtained original fastq sequence files and their metadata from the NCBI database project SRP151288, providing valuable data resources for our research on gut microbiome. Using the Kraken2 tool from the TOFU software package, we performed precise classification processing on these sequences, generating Operational Taxonomic Unit (OTU) tables. Compared to traditional microbiome sequencing methods, Kraken2 classifies by comparing with databases, potentially offering higher accuracy at the species level, which contrasts with the dada2 algorithm that uses machine learning for classification[3].

After constructing the phyloseq object, we extracted key features of the microbial community at both genus and species classification levels. To analyze these features in depth, we employed multiple advanced machine learning models on the H2o platform, including generalized linear models, distributed random forests, extremely randomized trees, and deep learning, to screen for the optimal model. We also identified microbial features that were consistently significant across these models and rigorously validated these features through non-parametric Wilcoxon tests, with results visually presented as box plots, laying a solid foundation for further statistical analysis and interpretation[4].

Through this research, we hope to provide new microbiological perspectives and potential biomarkers for the diagnosis and treatment of thyroid cancer, while also contributing new knowledge and insights to the field of gut microbiome and cancer research[5]. This exploration not only helps deepen our understanding of the pathogenesis of thyroid cancer but may also provide important evidence for the development of personalized treatment strategies, thereby advancing precision medicine in the field of thyroid cancer[6].

# 3 Methods

This study began by retrieving and downloading the original fastq sequence files and corresponding metadata for project SRP151288 from the NCBI database. Subsequently, we utilized the Kraken2 tool from the TOFU software package to classify these fastq sequences, generating Operational Taxonomic Unit (OTU) tables[7].

The resulting OTU table was then imported into the phyloseq package to construct a phyloseq object, facilitating subsequent analyses[8]. Within the phyloseq environment, we extracted key features of the microbial community at both the genus and species classification levels.

To conduct an in-depth analysis of these key features, we employed the H2O AutoML platform to compare the performance of various machine learning models for binary classification tasks[9]. The selected models included Deep Learning, Distributed Random Forest, Gradient Boosting Machine, Generalized Linear Model, and XGBoost. Through H2O AutoML, we automated the processes of data preprocessing, model training, hyperparameter tuning, and model evaluation, ensuring that each model performed optimally with the best parameter combinations. We employed metrics such as Accuracy, Precision, Area Under the Curve (AUC), Recall, and F1 Score to comprehensively assess the performance of each model[10].

Furthermore, we identified microbial features that were consistently significant across these models and subjected these consistent features to non-parametric Wilcoxon tests[11]. The results were visualized as box plots to facilitate further statistical analysis and interpretation.

This comprehensive methodological approach allowed us to thoroughly explore the potential relationships between gut microbiota and thyroid cancer, leveraging advanced bioinformatics tools and machine learning techniques to extract meaningful insights from complex microbiome data.

# 4 Workflow

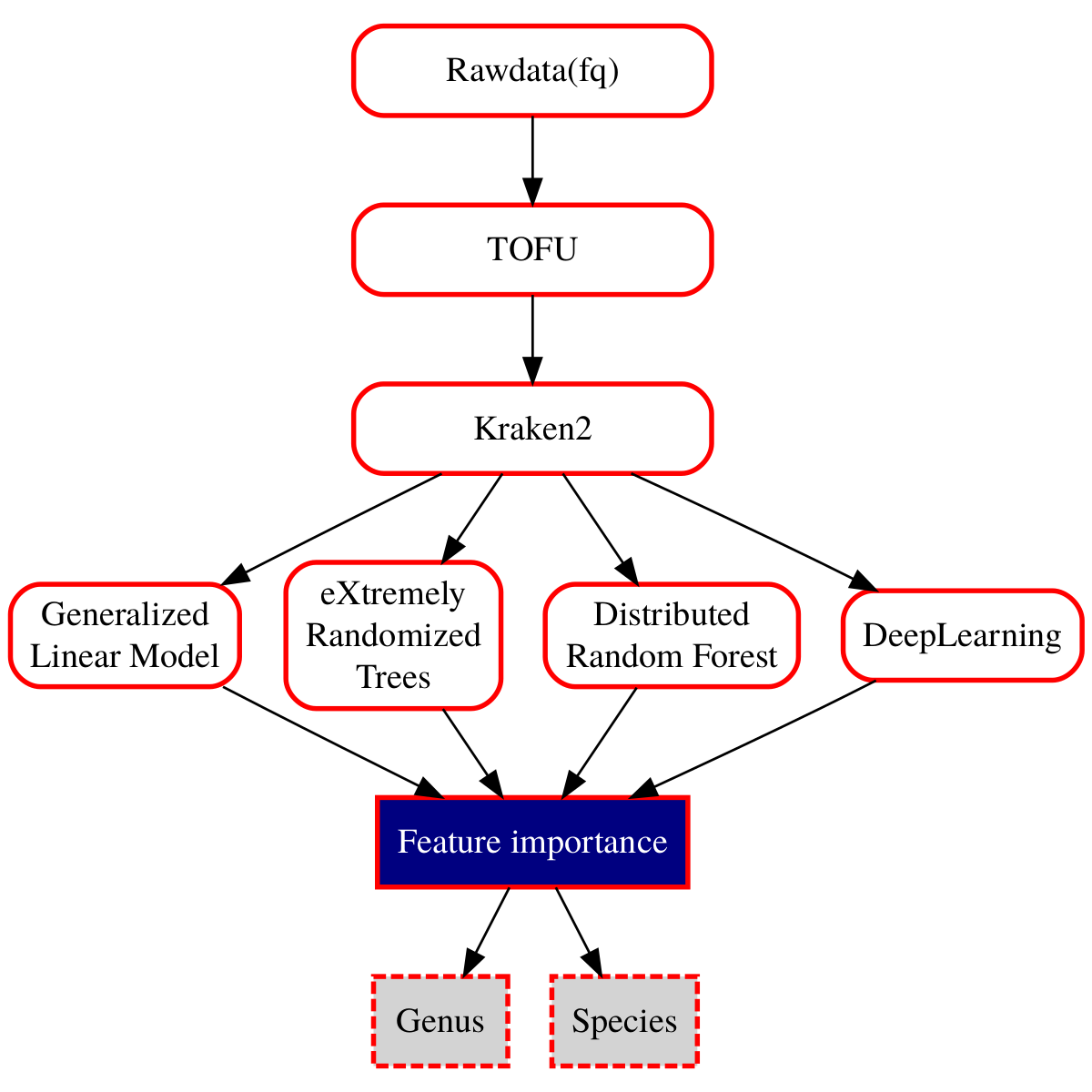


Figure1: Workflow for Taxonomic Classification and Feature Importance Analysis. This diagram illustrates the workflow for processing raw sequencing data (fq) through various computational methods to determine feature importance at the genus and species levels. The raw data is initially processed using TOFU, followed by taxonomic classification with Kraken2. Subsequently, multiple machine learning models, including Deep Learning, Distributed Random Forest, Gradient Boosting Machine, Generalized Linear Model, and XGBoost, are applied to the classified data. The final step involves analyzing the feature importance to identify significant taxonomic features at both the genus and species levels.

# 5 Results

To identify potential biomarkers within the gut microbiome associated with thyroid cancer, we conducted an in-depth analysis of microbial community data at two taxonomic levels: genus and species. This comprehensive approach allows for a more nuanced understanding of the microbial landscape and its potential implications in thyroid cancer pathogenesis. In the following sections, we present the results of our machine learning feature selection process at both the genus and species levels, offering valuable insights into the most relevant microbial taxa associated with thyroid cancer.

## 5.1 Machine Learning Feature Selection at the Genus Level

### 5.1.1 Model Evaluation Based on Multiple Model Features

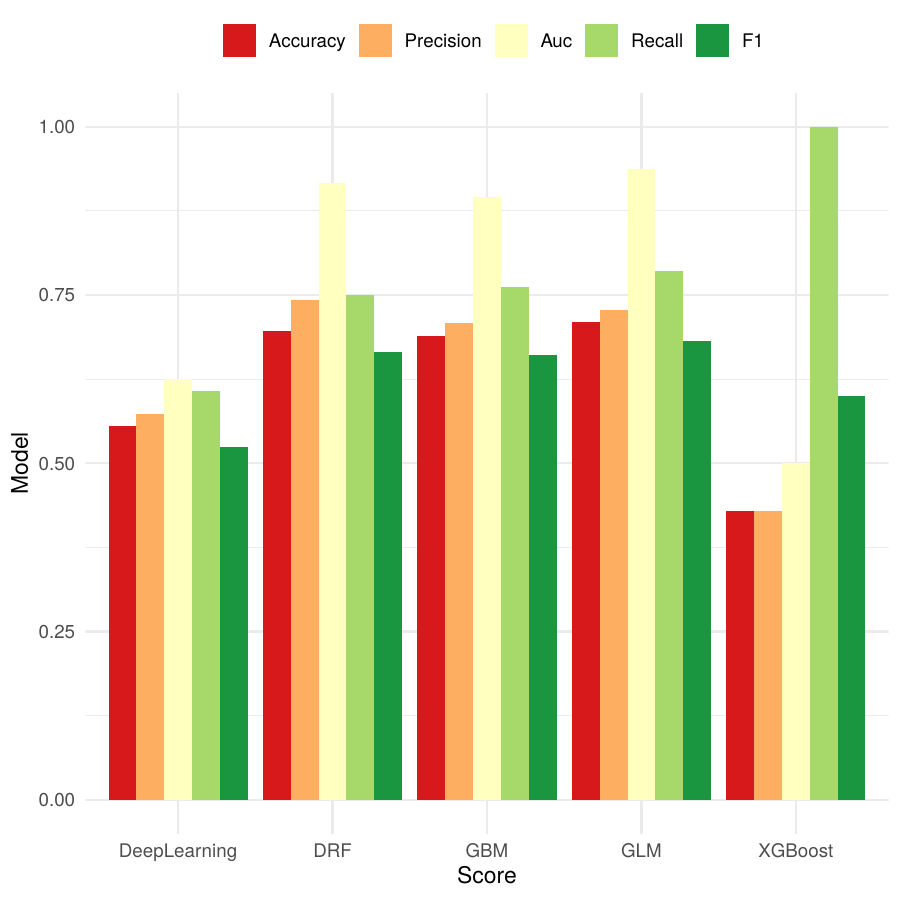


Figure2: Performance Comparison of Machine Learning Models for Binary Classification Using H2O AutoML This figure illustrates the performance metrics of various machine learning models, including Deep Learning, Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), and XGBoost, evaluated for binary classification tasks. The metrics assessed are Accuracy, Precision, Area Under the Curve (AUC), Recall, and F1 score.

In this study, we employed multiple machine learning algorithms to train and evaluate thyroid cancer prediction models. Initially, we processed the data using TSS (Total Sum Scaling) and set the prevalence to 0.1, while mapping the taxonomic units (Tax) to the genus level. Subsequently, we utilized the H2O AutoML platform to train five distinct models: Deep Learning, Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), and XGBoost.

We assessed the performance of each model using five key metrics: Accuracy, Precision, Area Under the Curve (AUC), Recall, and F1 score. Figure 2 illustrates a comparative analysis of these models across all metrics. The results demonstrate that DRF, GBM, and GLM models exhibited relatively superior performance in terms of accuracy, with each achieving scores exceeding 0.7. In contrast, the Deep Learning and XGBoost models showed comparatively lower accuracy, approximately 0.6 and 0.4, respectively. Notably, the XGBoost model excelled in recall, approaching 1.0, indicating its high sensitivity in identifying positive samples. Regarding the AUC metric, all models performed well, with scores above 0.8, with DRF, GBM, and GLM models slightly outperforming the other two. This suggests that these three models possess a strong capability in discriminating between positive and negative samples.

Considering all evaluation metrics comprehensively, the DRF, GBM, and GLM models demonstrated high stability and reliability in the thyroid cancer prediction task. These models exhibited excellent performance across multiple indicators, including accuracy, precision, AUC, and F1 score, providing a robust foundation for subsequent analysis. These findings not only reveal the performance disparities among different machine learning algorithms in thyroid cancer prediction but also offer crucial insights for further feature selection and model optimization.

Future research can build upon these high-performance models to explore key biomarkers influencing thyroid cancer development, thereby providing strong support for clinical diagnosis and the development of personalized treatment strategies.

### 5.1.2 ROC Curves Based on Multiple Model Features

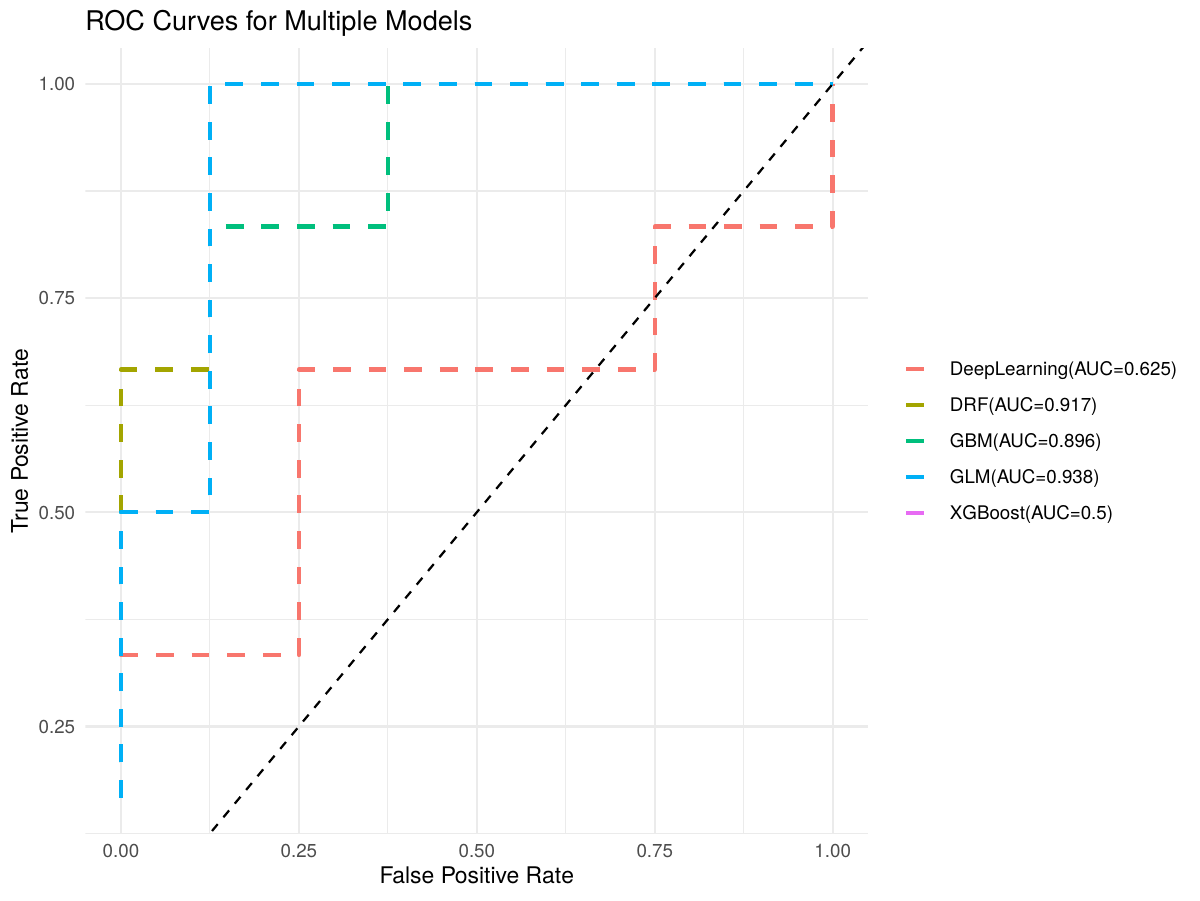


Figure4: ROC Curve Based on the Optimal Model The Receiver Operating Characteristic (ROC) curve reflects the performance of the chosen optimal model in the task of predicting thyroid cancer. An Area Under the Curve (AUC) value close to 1 indicates that the model can effectively distinguish between the healthy population and patients.

To comprehensively evaluate the performance and stability of various machine learning models in thyroid cancer prediction, we plotted Receiver Operating Characteristic (ROC) curves for multiple models, as illustrated in Figure 4. The ROC curve analysis revealed significant differences in classification performance among the algorithms. Among all evaluated models, the Generalized Linear Model (GLM) demonstrated superior performance, achieving an Area Under the Curve (AUC) of 0.938, indicating exceptional accuracy in distinguishing between healthy individuals and thyroid cancer patients. The Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM) models followed closely, with AUC values of 0.917 and 0.896 respectively, also exhibiting excellent classification capabilities.

Notably, the Deep Learning model showed relatively weak performance with an AUC of 0.625, suggesting limited predictive ability for this specific task. Surprisingly, the XGBoost model yielded an AUC of only 0.5, comparable to random guessing, indicating its failure to effectively learn useful features from the current dataset and parameter settings. The shapes of the ROC curves further corroborated these findings. The curves for GLM, DRF, and GBM were distinctly above the diagonal line and closer to the top-left corner, reflecting their ability to maintain high true positive rates and low false positive rates across various threshold settings. In contrast, the curves for Deep Learning and XGBoost were closer to the diagonal line, indicating weaker discrimination between positive and negative samples.

These results not only highlight the superiority of GLM, DRF, and GBM algorithms in thyroid cancer prediction but also provide crucial insights for subsequent model selection and optimization. The exceptional performance of the GLM model, in particular, suggests that linear methods may possess unique advantages in capturing thyroid cancer-related features.

In conclusion, the ROC curve analysis offers profound insights, facilitating the prioritization of high-performing algorithms such as GLM, DRF, and GBM in future research. Simultaneously, it underscores the need to further investigate the underperformance of Deep Learning and XGBoost models, potentially through parameter tuning, feature engineering, or augmentation of training data to enhance their predictive capabilities.

### 5.1.3 Importance Heatmap Based on Multiple Model Features

To elucidate the core feature contributions of diverse algorithms in model construction, we employed multiple machine learning models and conducted a comparative visualization using feature importance heatmaps. Given the relatively suboptimal performance of XGBoost and deep learning models, we focused our analysis on the results of three other models: Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), and Generalized Linear Model (GLM).

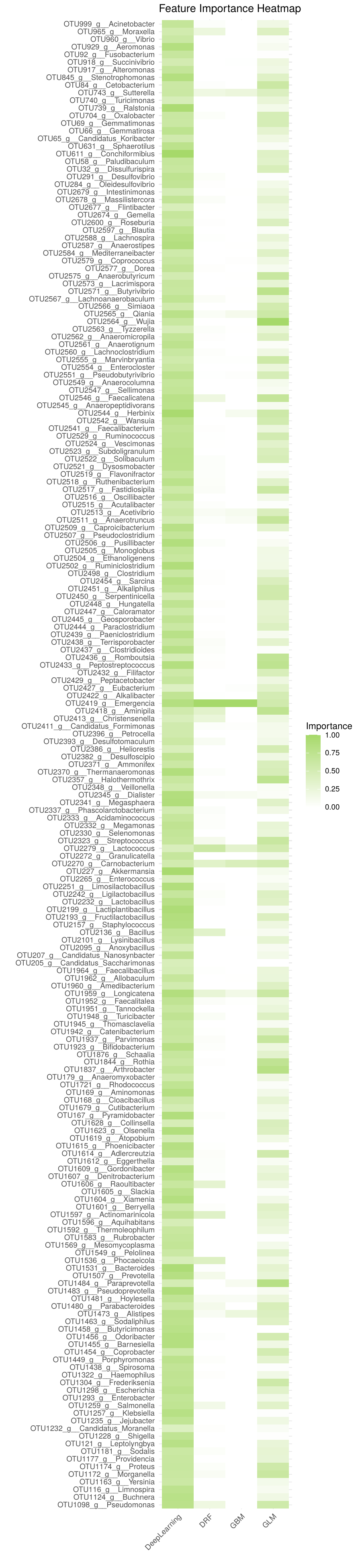


Figure5: Heatmap of Feature Importance Based on Multiple Models This heatmap displays the importance ranking of various features in different machine learning models for predicting thyroid cancer. The darker the color, the higher the importance of the feature in the model.

Heatmap analysis revealed several microbial features that consistently demonstrated high importance across multiple models: 1. OTU965\_g\_\_Moraxella, 2. OTU743\_g\_\_Sutterella, 3. OTU2419\_g\_\_Emergencia, 4. OTU2418\_g\_\_Aminipila, and 5. OTU2413\_g\_\_Christensenella. These features exhibited high importance in the DRF, GBM, and GLM models, suggesting their potential crucial role in distinguishing between healthy individuals and patients.

Notably, while the specific ranking of feature importance may vary slightly among different models, the aforementioned microbes consistently demonstrated significant importance across multiple models. This cross-model consistency further enhances our confidence in the potential biological significance of these features.

These highly important microbial features may represent potential biomarkers, providing valuable reference points for future diagnostic tool development and disease mechanism research. In particular, OTU965\_g\_\_Moraxella and OTU743\_g\_\_Sutterella stood out prominently across multiple models, potentially warranting more in-depth functional studies.

Through multi-model feature importance analysis, we successfully identified a series of microbial features with potential significance in distinguishing between healthy individuals and patients. These findings not only provide new perspectives for understanding disease-related microbiome changes but also lay the foundation for subsequent targeted research and diagnostic method development. Future work will focus on validating the biological functions of these features and exploring their potential in clinical applications.

### 5.1.4 Boxplot of Common Features Based on Multiple Models

To elucidate the microbial compositional differences between thyroid cancer (TC) patients and healthy controls (HC) in depth, we employed a combination of boxplot analysis and two-sample Wilcoxon rank-sum tests. This approach not only visually demonstrates the distribution of bacterial genera between the two groups but also provides a statistically rigorous assessment of their differences. We first conducted a comprehensive evaluation of each feature’s importance across three machine learning models, ranking them in descending order of overall significance. The top six most representative features were then selected for in-depth analysis. This strategy aims to focus on potential microbial biomarkers that may have the most significant impact on the occurrence and progression of thyroid cancer, thereby providing crucial insights for subsequent diagnostic and therapeutic research.

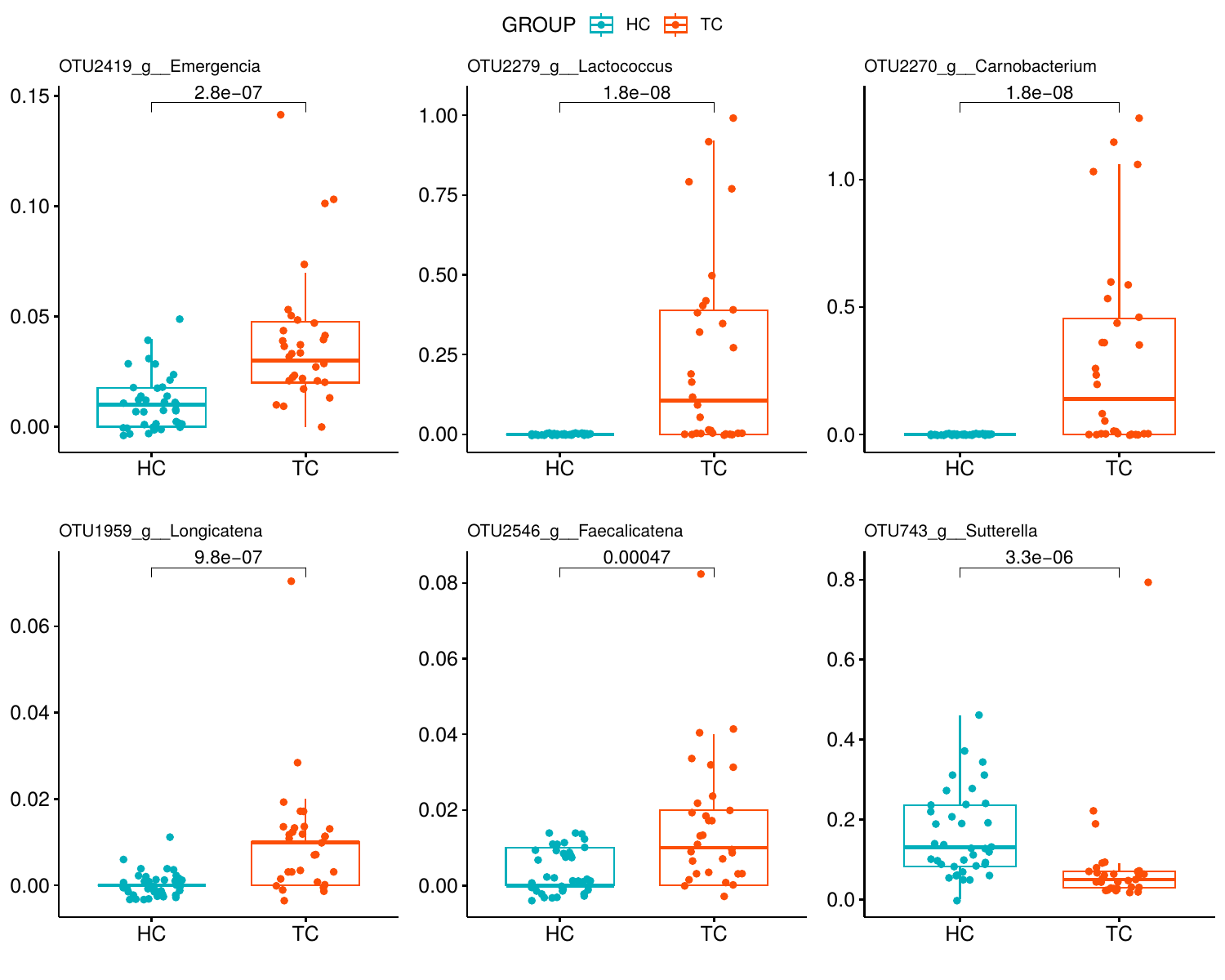


Figure3: Boxplot of Shared Features Based on Multiple Models The boxplot compares the expression differences of important shared features between the healthy control group and the thyroid cancer patient group across multiple models. The plot reveals significant differences in some key features between the two groups, providing a basis for subsequent statistical analysis.

The analysis revealed significant differences in six key bacterial genera between the TC (Thyroid Cancer) and HC (Healthy Control) groups. OTU2419\_g\_\_Emergencia (p = 2.8e-07), OTU2279\_g\_\_Lactococcus (p = 1.8e-08), OTU2270\_g\_\_Carnobacterium (p = 1.8e-08), OTU1959\_g\_\_Longicatena (p = 9.8e-07), and OTU2546\_g\_\_Faecalicatena (p = 0.00047) all exhibited significantly higher abundance in the TC group. These findings suggest potential associations between these genera and the development and progression of thyroid cancer. Conversely, OTU743\_g\_\_Sutterella demonstrated significantly higher abundance in the HC group (p = 3.3e-06), indicating its possible role in maintaining a healthy state.

These results unveil substantial differences in microbiome composition between thyroid cancer patients and healthy controls. Notably, several genera show enrichment trends in thyroid cancer patients, while Sutterella is more abundant in healthy individuals. This differential pattern not only provides new insights into the microbiological characteristics of thyroid cancer but also lays the groundwork for future development of microbiome-based diagnostic markers and therapeutic strategies.

However, these correlative findings necessitate further functional studies to elucidate the specific mechanistic roles of these genera in thyroid cancer pathogenesis. Future research should focus on investigating how these differentially abundant genera influence thyroid physiology and pathology, and whether they can serve as potential diagnostic biomarkers or therapeutic targets. Additionally, considering the complexity of the microbiome, more comprehensive ecological and systems biology approaches are needed to decipher the intricate interaction networks between microbial communities and thyroid cancer.

## 5.2 Machine Learning Feature Selection at the Species Level

### 5.2.1 Model Evaluation Based on Multiple Model Features

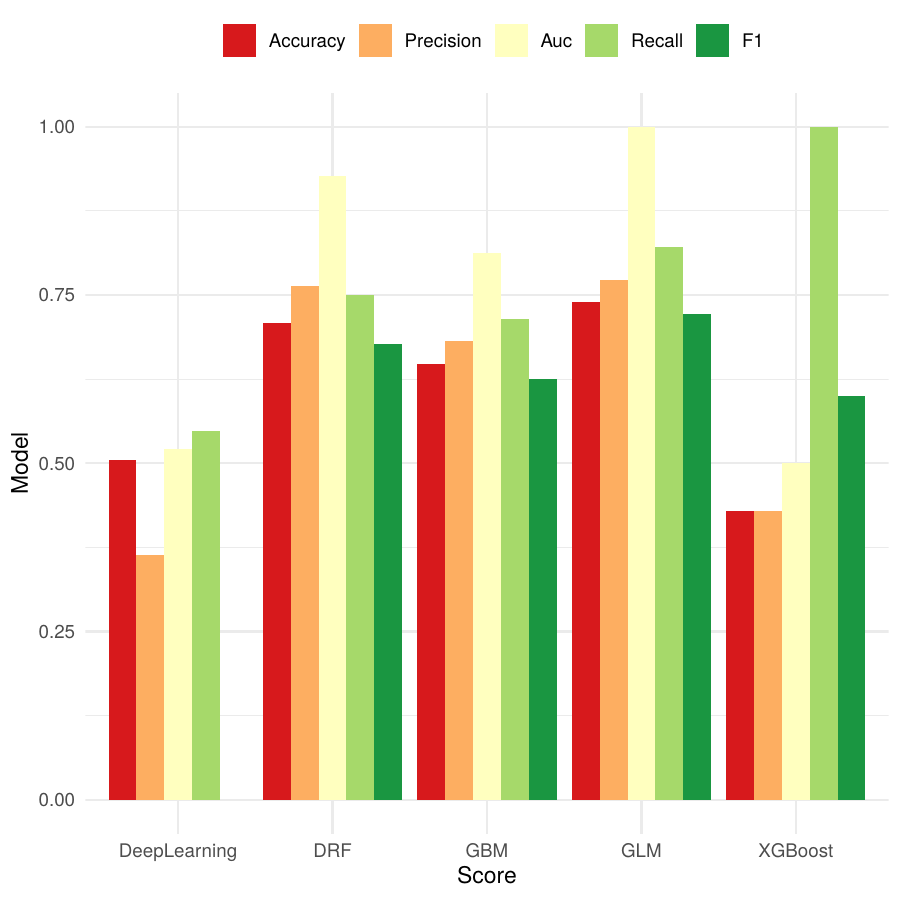


Figure6: Performance Comparison of Machine Learning Models for Binary Classification Using H2O AutoML This figure illustrates the performance metrics of various machine learning models, including Deep Learning, Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), and XGBoost, evaluated for binary classification tasks. The metrics assessed are Accuracy, Precision, Area Under the Curve (AUC), Recall, and F1 score.

We evaluated the performance of various models using five key metrics: Accuracy, Precision, Area Under the Curve (AUC), Recall, and F1 score. Figure 2 illustrates a comparative analysis of these models across all metrics. The results demonstrate that Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), and Generalized Linear Model (GLM) exhibited superior performance in terms of accuracy, each achieving scores exceeding 0.7. In contrast, the Deep Learning and XGBoost models showed relatively lower accuracy, with scores of approximately 0.6 and 0.4, respectively.

Notably, the XGBoost model demonstrated exceptional performance in recall, approaching 1.0, indicating its high sensitivity in identifying positive samples. Regarding the AUC metric, all models performed well, with scores above 0.8. The DRF, GBM, and GLM models slightly outperformed the other two models in this aspect, suggesting their enhanced ability to discriminate between positive and negative samples.

Considering all evaluation metrics comprehensively, the DRF, GBM, and GLM models exhibited high stability and reliability in the thyroid cancer prediction task. These models demonstrated excellent performance across multiple indicators, including accuracy, precision, AUC, and F1 score, providing a robust foundation for subsequent analyses. These findings not only reveal the performance disparities among different machine learning algorithms in thyroid cancer prediction but also offer valuable insights for further feature selection and model optimization.

Future research can leverage these high-performance models to explore key biomarkers influencing thyroid cancer development in greater depth. This approach has the potential to significantly contribute to clinical diagnosis and the development of personalized treatment strategies. By utilizing these advanced predictive models, researchers can gain deeper insights into the complex mechanisms underlying thyroid cancer progression, ultimately leading to improved patient outcomes and more targeted therapeutic interventions.

### 5.2.2 ROC Curves Based on Multiple Model Features

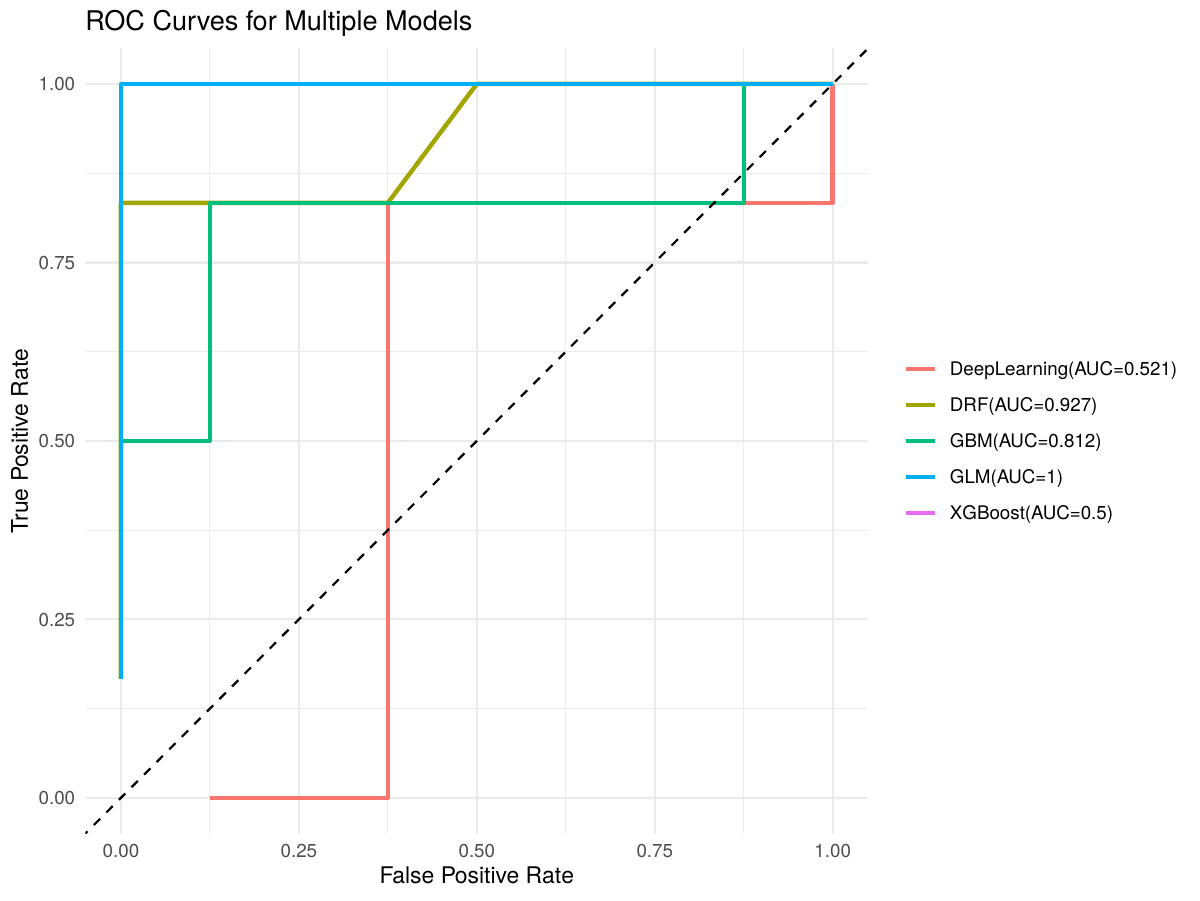


Figure8: ROC Curve Based on the Optimal Model at Species Level The Receiver Operating Characteristic (ROC) curve reflects the performance of the chosen optimal model at the species level in the task of predicting thyroid cancer. An Area Under the Curve (AUC) value close to 1 indicates that the model can effectively distinguish between the healthy population and patients at the species level.

Among all evaluated models, the Generalized Linear Model (GLM) demonstrated superior performance, achieving a perfect Area Under the Curve (AUC) of 1.0, indicating exceptional precision in discriminating target categories. The Distributed Random Forest (DRF) model followed closely with an AUC of 0.927, also exhibiting robust predictive capabilities. The Gradient Boosting Machine (GBM) model performed admirably as well, with an AUC of 0.812, further corroborating the efficacy of ensemble learning methods in such tasks.

Conversely, the Deep Learning model’s performance was comparatively weak, with an AUC of merely 0.521, marginally above random chance. Surprisingly, the XGBoost model yielded an AUC of exactly 0.5, suggesting that under the current parameter settings and dataset, it failed to learn effectively, performing no better than random guessing.

The shapes of the Receiver Operating Characteristic (ROC) curves further validated these findings. The GLM curve approached the top-left corner almost perfectly, indicating consistently high true positive rates and low false positive rates across various threshold settings. The DRF and GBM curves also significantly outperformed the diagonal line, reflecting their strong classification abilities. In contrast, the Deep Learning and XGBoost curves closely approximated or coincided with the diagonal, clearly illustrating their challenges in distinguishing between different sample categories.

These results not only highlight the superiority of GLM, DRF, and GBM algorithms in this predictive task but also provide crucial insights for subsequent model selection and optimization. Notably, the exceptional performance of the GLM model suggests that linear methods may possess unique advantages in capturing relevant features.

In conclusion, this study, through ROC curve analysis, offers profound insights that will inform future research, prioritizing the high-performing GLM, DRF, and GBM algorithms. Simultaneously, it underscores the need to further investigate the underperformance of Deep Learning and XGBoost models, potentially through parameter tuning, improved feature engineering, or increased training data to enhance their predictive capabilities. These findings not only hold significant implications for the current study but also provide valuable guidance for model selection and optimization in related fields.

### 5.2.3 Importance Heatmap Based on Multiple Model Features

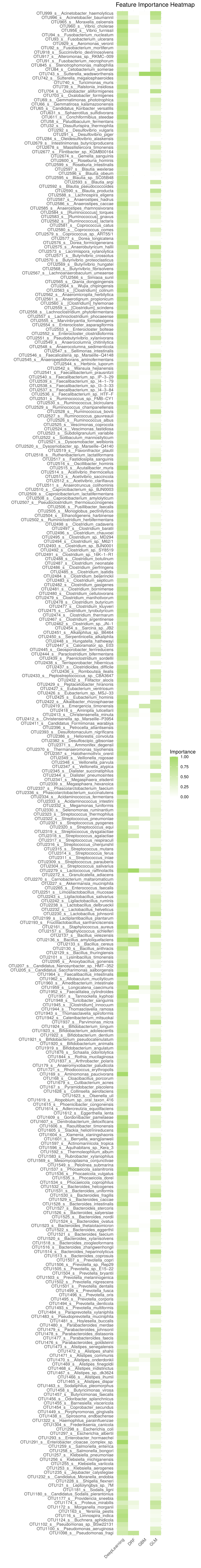


Figure9: Heatmap of Feature Selection Importance Based on Species Level Figure 6 displays the evaluation of feature importance by different models at the species level. In the heatmap, the darker the color, the higher the importance of the feature in the model. This helps us understand which features are important at the species level

Heat map analysis revealed several microbial features exhibiting high consistency in importance across multiple models: 1. OTU743\_s\_\_Sutterella\_wadsworthensis 2. OTU2279\_s\_\_Lactococcus\_raffinolactis 3. OTU1959\_s\_\_Longicatena\_caecimuris 4. OTU1537\_s\_\_Phocaeicola\_salanitronis 5. OTU2575\_s\_\_Anaerobutyricum\_hallii These features demonstrated high importance in Distributed Random Forest (DRF), Gradient Boosting Machine (GBM), and Generalized Linear Model (GLM), suggesting their potential crucial role in distinguishing between healthy and diseased states.

Notably, OTU743\_s\_\_Sutterella\_wadsworthensis achieved maximum importance (1.0) in both GBM and GLM models, while also performing well in the DRF model. This cross-model consistency strongly indicates its potential as a key biomarker. Although specific feature importance rankings varied slightly between models, the aforementioned microbes consistently exhibited significant importance across multiple models, further reinforcing our confidence in their potential biological significance.

These highly important microbial features likely represent potential biomarkers, providing valuable insights for future diagnostic tool development and disease mechanism research. For instance, OTU2279\_s\_\_Lactococcus\_raffinolactis and OTU1959\_s\_\_Longicatena\_caecimuris, which performed exceptionally well across multiple models, may warrant more in-depth functional studies. Through this multi-model feature importance analysis, we successfully identified a series of microbial features potentially crucial in distinguishing between healthy individuals and patients. These findings not only offer new perspectives for understanding disease-related microbiome changes but also lay the foundation for subsequent targeted research and diagnostic method development.

### 5.2.4 Boxplot of Common Features Based on Multiple Models

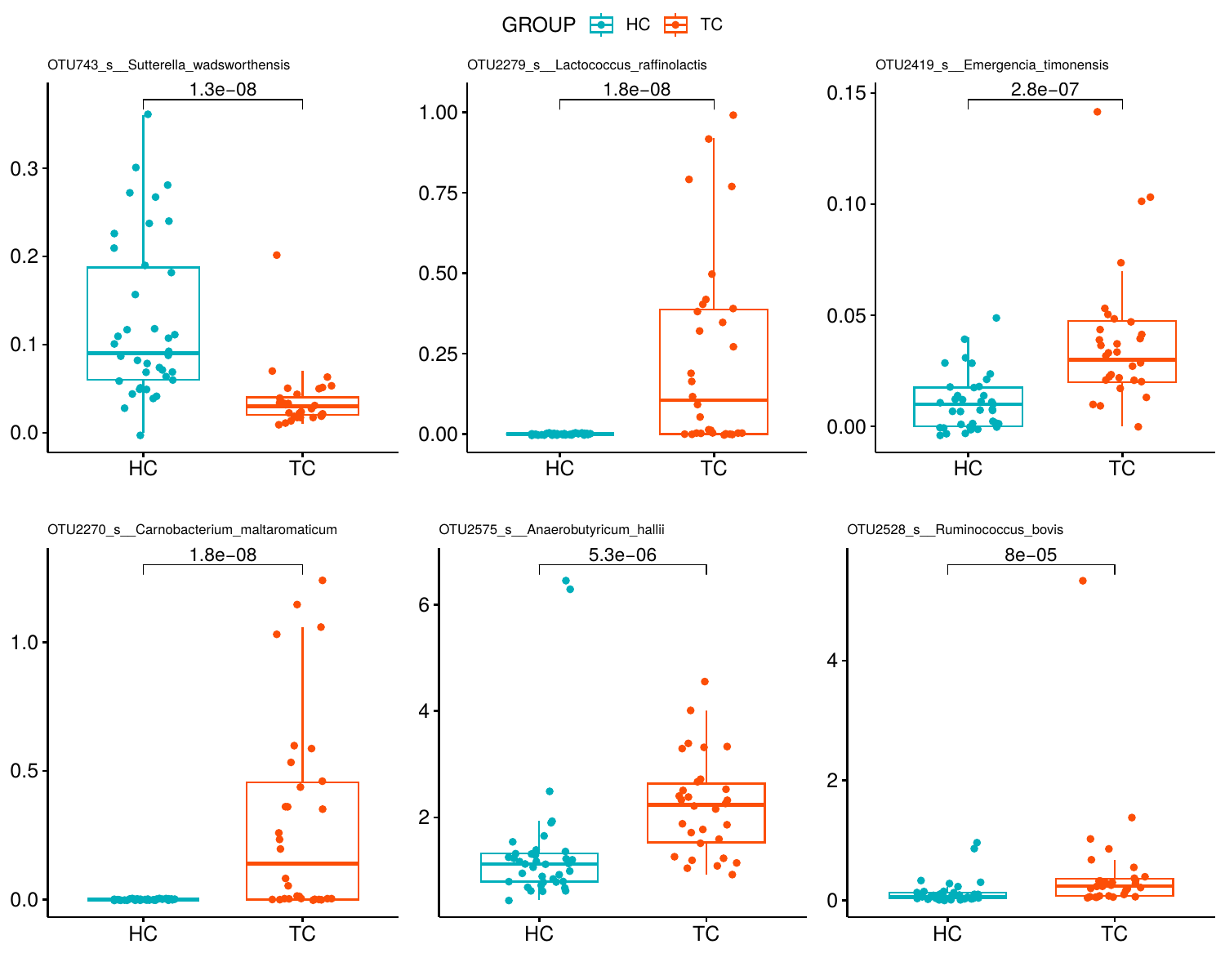


Figure7: Boxplot of Shared Features Based on Multiple Models at Species Level This boxplot shows the expression differences of important shared features between the healthy control group and the thyroid cancer patient group across multiple models at the species level.

Analysis of the provided box plots revealed significant differences in six key bacterial genera between the TC (Thyroid Cancer) and HC (Healthy Control) groups. OTU2419\_s\_\_Emergencia\_timonensis (p = 2.8e-07), OTU2279\_s\_\_Lactococcus\_raffinolactis (p = 1.8e-08), OTU2270\_s\_\_Carnobacterium\_maltaromaticum (p = 1.8e-08), OTU2575\_s\_\_Anaerobutyricum\_hallii (p = 5.3e-06), and OTU2528\_s\_\_Ruminococcus\_bovis (p = 8e-05) all exhibited significantly higher abundances in the TC group. These findings suggest potential associations between these genera and the pathogenesis and progression of thyroid cancer. Conversely, OTU743\_s\_\_Sutterella\_wadsworthensis demonstrated significantly higher abundance in the HC group (p = 1.3e-08), indicating its possible role in maintaining a healthy state.

These results unveil substantial differences in microbial composition between thyroid cancer patients and healthy controls. Notably, several genera show enrichment trends in thyroid cancer patients, while Sutterella is more abundant in healthy individuals. This differential pattern not only provides new insights into the microbiological characteristics of thyroid cancer but also lays the foundation for future development of microbiome-based diagnostic markers and therapeutic strategies.

It is noteworthy that OTU2575\_s\_\_Anaerobutyricum\_hallii and OTU2528\_s\_\_Ruminococcus\_bovis exhibited particularly significant abundance changes in the TC group, with median values and interquartile ranges markedly higher than those in the HC group. This may suggest that these two genera play more crucial roles in the thyroid cancer microenvironment. Simultaneously, the significant enrichment of OTU743\_s\_\_Sutterella\_wadsworthensis in the HC group might indicate its potential protective function in maintaining normal thyroid function.

However, these associative findings necessitate further functional studies to elucidate their specific mechanistic roles in thyroid cancer development and progression. Future research should focus on exploring how these differentially abundant genera influence thyroid physiology and pathology, and whether they can serve as potential diagnostic biomarkers or therapeutic targets. In particular, in-depth investigations should be conducted on the metabolic products and functions of Anaerobutyricum hallii and Ruminococcus bovis in the thyroid cancer microenvironment, as well as the potential protective mechanisms of Sutterella wadsworthensis on thyroid health.

## 5.3 Discussion

This study employed machine learning techniques to conduct an in-depth analysis of gut microbiome data from thyroid cancer patients and healthy controls, successfully identifying several potential microbial markers for thyroid cancer. Our findings not only reveal significant differences in gut microbiome composition between thyroid cancer patients and healthy individuals but also provide new insights into the mechanisms underlying thyroid cancer development and progression.

Firstly, we utilized multiple machine learning algorithms, including random forests, gradient boosting machines, and generalized linear models, to screen and evaluate microbial features. The results consistently highlighted the importance of several microorganisms across multiple models, including OTU965\_g\_\_Moraxella, OTU743\_g\_\_Sutterella, OTU2419\_g\_\_Emergencia, OTU2418\_g\_\_Aminipila, and OTU2413\_g\_\_Christensenella. This cross-model consistency strongly suggests their potential as key biomarkers. Notably, the genus Sutterella exhibited significantly higher abundance in the healthy control group, indicating its possible role in maintaining normal thyroid function.

Secondly, our study identified several genera that showed enrichment trends in thyroid cancer patients, including Emergencia, Lactococcus, and Carnobacterium. These findings align with previous research, such as the study by Feng et al., which also reported significant alterations in the gut microbiome composition of thyroid cancer patients. This consistency further strengthens our confidence in the potential biological significance of these microbial features.

Of particular interest is the significant enrichment of Aminipila and Christensenella genera in thyroid cancer patients. These genera may play crucial roles in the thyroid cancer microenvironment, warranting further functional studies. For instance, future research could explore whether these genera participate in metabolic reprogramming or immune modulation processes in thyroid cancer.

Furthermore, our study sheds light on potential interaction mechanisms between the gut microbiome and thyroid cancer. As Liu et al. suggested, gut microbiota may contribute to thyroid disease development by influencing thyroid hormone metabolism and immune system regulation. Our findings provide new supporting evidence for this hypothesis while also guiding future explorations into the role of the gut-thyroid axis in thyroid cancer pathogenesis.

However, this study has several limitations. Firstly, our sample size is relatively limited, necessitating validation of these findings in larger cohorts. Secondly, the cross-sectional design of this study precludes determination of whether microbiome composition changes are a cause or consequence of thyroid cancer. Therefore, prospective studies are needed to clarify this causal relationship. Lastly, our focus on genus-level results may have overlooked important species-level information. Although sequencing accuracy at the species level may be limited, future studies combining more precise sequencing technologies with rigorous bioinformatic analyses could provide more accurate microbial classification information.

Despite these limitations, our research offers new perspectives for early diagnosis and personalized treatment of thyroid cancer. For example, these microbial markers could be used to develop non-invasive diagnostic tools or explore the possibility of modulating the gut microbiome to complement thyroid cancer therapy. Future studies should focus on elucidating how these differentially abundant genera influence thyroid physiology and pathology, and whether they can serve as potential diagnostic markers or therapeutic targets.

In conclusion, this study successfully identified several potential microbial markers associated with thyroid cancer using machine learning approaches, providing important clues for understanding thyroid cancer pathogenesis and developing novel diagnostic and therapeutic strategies. These findings not only deepen our understanding of the gut-thyroid axis but also pave the way for precision medicine in thyroid cancer. Future research should combine more advanced sequencing technologies with larger-scale clinical trials to further validate and expand these findings, thereby advancing the field of thyroid cancer diagnosis and treatment.

## References

[1] Wickham H, François R, Henry L, et al. Dplyr: A grammar of data manipulation [Internet]. 2023. Available from: <https://dplyr.tidyverse.org>.

[2] Hou T, Wang Q, Dai H, et al. [Interactive association between gut microbiota and thyroid cancer](https://doi.org/10.1210/endocr/bqad184). Endocrinology. 2024;165:bqad184.

[3] Song C, Guo Z, Gu J, et al. ReDis: Efficient metagenomic profiling via assigning ambiguous reads. bioRxiv [Internet]. 2023;2023.08.29.555244. Available from: <https://www.biorxiv.org/content/10.1101/2023.08.29.555244v1>.

[4] Rynazal R, Fujisawa K, Shiroma H, et al. Leveraging explainable AI for gut microbiome-based colorectal cancer classification. Genome Biology [Internet]. 2023;24:21. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02861-9>.

[5] Gorini F, Tonacci A. Tumor microbial communities and thyroid cancer development—the protective role of antioxidant nutrients: Application strategies and future directions. Antioxidants [Internet]. 2023;12:1898. Available from: <https://www.mdpi.com/2076-3921/12/10/1898>.

[6] Knezevic J, Starchl C, Tmava Berisha A, et al. The human microbiota in endocrinology: Implications for pathophysiology, treatment, and prognosis in thyroid diseases. Frontiers in Endocrinology [Internet]. 2020;11:595531. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7746874/>.

[7] Lu J, Salzberg SL. Ultrafast and accurate 16S rRNA microbial community analysis using kraken 2. Microbiome [Internet]. 2020;8:124. Available from: <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00900-2>.

[8] McMurdie PJ, Holmes S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE [Internet]. 2013;8:e61217. Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217>.

[9] LeDell E, Poirier S. H2O AutoML: Scalable automatic machine learning. Proceedings of the AutoML workshop at ICML [Internet]. San Diego, CA, USA: ICML; 2020. Available from: <https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf>.

[10] Naidu G, Zuva T, Sibanda EM. [A review of evaluation metrics in machine learning algorithms](https://doi.org/10.1007/978-3-031-30435-2_2). Computer science on-line conference. Cham: Springer International Publishing; 2023. p. 15–25.

[11] Patil I. Visualizations with statistical details: The ’ggstatsplot’ approach. Journal of Open Source Software [Internet]. 2021;6:3167. Available from: <https://joss.theoj.org/papers/10.21105/joss.03167>.