A Supervised Learning Approach to Classify Clinical Outcomes in *Helicobacter pylori*-positive Patients

Venkatesh Narasimhan[1,*], Sreya Pulakkat Warrier[2,*], Jobin Jacob John[1], Monisha Priya T[1], Niriksha Varadaraj[1], Greeshma Grace Thomas[3], Balaji Veeraraghavan[1,**]

[1] Department of Clinical Microbiology, Christian Medical College and Hospital, Vellore, India

[2] School of Biosciences and Technology, Vellore Institute of Technology, Vellore, India

[3] Department of General Medicine, Christian Medical College and Hospital, Vellore, India

[*] Authors contributed equally

[**] Corresponding author

Correspondence:

Prof. Balaji Veeraraghavan

Department of Clinical Microbiology,

Christian Medical College,

Vellore,

India

Email: vbalaji@cmcvellore.ac.in

Abstract

Background: *Helicobacter pylori* (*H. pylori*) is a gastric pathogen with global prevalence that is implicated in a lot of clinical outcomes, ranging from gastritis to gastric cancer. While chronic gastritis is a nearly universal outcome among infected individuals, only a small subset progresses to severe diseases like gastric cancer, with these outcomes being governed by a mixture of bacterial, host and environmental factors. This study presents a supervised learning framework in order to predict whether an *H. pylori* infection will result in gastric cancer or a non-malignant outcome with a combination of clinical and genome-derived features.

Methods: A dataset with 1,363 *H. pylori* genomes containing host metadata and annotated genomic features was curated. Feature extraction was performed using gene presence/absence profiles, various sequence descriptors, and variant-level annotation features. A white-box logistic regression model and black-box eXtreme Gradient Boosting (XGBoost) classifier and Random Forest models were trained and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) was used for addressing class imbalance. SHapley Additive exPlanations (SHAP) values were used to assess model interpretability.

Results: The baseline model showed an Area Under Receiver Operating Characteristic curve (AUROC) of 0.89 while the XGBoost model and Random Forest models displayed an AUROC of 0.95. The black-box models improved recall for the gastric cancer group with 0.83 for XGBoost and 0.82 for Random Forest while the baseline model had a recall of 0.73. Feature importance analyses were also used to identify the key drivers of gastric cancer prediction including host age, geographic origin and specific disruptive mutations. These findings support

the use of machine learning in biomarker-guided decision-making in assessing gastric disease risk.

Conclusions: The black-box models demonstrated improved performance with a recall value greater than 0.8 compared to the baseline model. Age was implicated as the top predictor in gastric cancer development across all models, along with a variety of genomic-level features. Further analysis of contributing features to check for causation could expand current research on the links between *H. pylori* and gastric cancer.

Keywords: *Helicobacter pylori, Gastric cancer prediction, Supervised machine learning, Genomic sequence-based feature extraction, SMOTE-NC, Ensemble learning, SHAP interpretability*

1. Background:

Helicobacter pylori (*H. pylori*) is a highly motile, helical and gram-negative bacterium that is known for colonising the human gastrointestinal tract, and primarily associated with chronic active gastritis. *H. pylori* infection is strongly linked to the development of serious gastrointestinal diseases and accounts for up to 90% of duodenal ulcers, and around 80% of stomach ulcers and 90% of gastric carcinoma and mucosa-associated lymphoid tissue (MALT) lymphoma cases[1]. In accordance with the burden posed by these diseases, the World Health Organisation (WHO) recommended prioritising *H. pylori* eradication strategies in 2014 to help in reducing stomach cancer mortality rates worldwide.

At the global level, *H. pylori* affects almost half of the world's population, and has a disproportionately high prevalence of around 50.8% in resource-limited regions, which are often characterized by suboptimal sanitation and healthcare access [2]. Gastric cancer is also relatively uncommon among younger individuals with fewer than 10% of cases occurring before the age of 45. There is a sharp spike in the lifetime risk of gastric cancer for infected individuals, with the probability being nearly double compared to uninfected individuals[3].

The clinical outcomes of *H. pylori* infection considerably vary across geographic populations and are influenced by a variety of factors like bacterial genotype diversity, host immune responses, and environmental factors. The bacterium's ability to thrive in the acidic gastric environment has been linked to specialized virulence mechanisms. These help it facilitate persistent colonization and chronic inflammation, which dictate the host's gastritis phenotype and disease trajectory. There are several *H. pylori* genes that have been implicated in the development of severe gastroduodenal diseases. The cytotoxin-associated gene A (*cagA)* gene, encoding the CagA protein is found at the end of the cag pathogenicity island (cagPAI), and is associated with increased cancer risk and severe infections[4]. The marker is more discriminative in Western countries on account of only around 60% of strains carrying *cagA*, compared to over 90% of East Asian strains being *cagA*-positive[5]. The vacuolating cytotoxin A (*vacA)* gene encodes the vacA cytotoxin which helps in inducing vacuole formation, autophagy modulation and evading immune responses[6,7]. Strains harbouring the *vacA* s1/m1/*cagA*+ genotype show a 4.8-fold increased risk of progressing to precancerous lesions compared to *vacA* s2/m2/*cagA*- strains[8]. The *babA2* gene, encoding the blood group antigen binding adhesin (BabA) that binds to mucosal Lewis[b] blood group antigens, helps in aiding colonisation and increasing bacterial load[9]. The triple positive genotype possessing *babA2, vacA* s1 and *cagA* genes offer stronger gastric

cancer discrimination than *vacA* and *cagA* alone[10]. The outer inflammatory protein A (*oipA*) gene promotes epithelial attachment, IL-8-mediated inflammation and disruption of cell turnover, and is often found co-expressed with other virulence genes[11–16]. *homB* gene has been found to aid adhesion and IL-8 induction, while *sabA* promotes neutrophil activation and inflammation through binding to sialylated Lewis antigens (sLe^X)[17–20].

The *htrA* gene is closely linked to tumor suppression and metastasis, and the *iceA1* gene has also been significantly associated with gastric cancer[21–23]. The *hopQ* type I allele is linked with increased gastric cancer and gastric ulcer risk, especially in East Asia where it is commonly present with *cagA*, while type II *hopQ* is more frequent in Western strains[24].

Diagnosing *H. pylori* infection is reliant on a combination of methods. These include invasive methods such as endoscopy and biopsy, and non-invasive approaches like urea breath tests, stool antigen tests, and serological assays. The diagnostic tool choice is guided by a lot of factors like patient history, clinical presentation and resource availability. Typical eradication therapy usually involves a potent acid suppressant with antibiotics and bismuth compounds in some cases. Curbing the alarming rise of antibiotic resistance among *H. pylori* strains relies on judicious antibiotic use and resistance surveillance methods[25].

Understanding the genetic diversity of *H. pylori* and its associated phenotypic variability can be used to build accurate prediction models using machine learning approaches. This could help improve therapeutic approaches by stratifying patients based on infection outcomes[26].

Despite there being some studies on predicting gastric cancer in *H. pylori* infected patients, they have primarily relied on clinical and lifestyle data and often included patients without confirmed

H. pylori infection[27–29]. To our knowledge, there have been no studies that have integrated clinical host metadata, aggregated variant annotation features from VCF files, virulence gene presence and absence data, and sequence-derived features from *H. pylori* genomes in a single, unified machine learning framework. In this study, we focused solely on *H. pylori* infected patients and developed a machine learning pipeline to classify gastric cancer versus non-gastric cancer cases using a combination of the above mentioned features. A comprehensive summary of the study workflow is provided in Figure 1.

2. Methods

2.1 Data Collection

Helicobacter pylori genomes from NCBI and Enterobase were filtered for the presence of host-age, host-sex, host geographical location and disease phenotype information, resulting in 1587 genomes. The genomes from Enterobase were pre-assembled and case controls were excluded.

NCBI Biosample Metadata was similarly downloaded and case controls and duplicates were excluded. Pre-assembled genomes from GenBank and RefSeq were downloaded using NCBI Datasets, while Sequence Read Archive (SRA) database sequences consisting of high-throughput short reads were downloaded through Fastq-dump from SRA Toolkit (version 3.2.1) .

2.2 Genome Quality Control and Assembly

The unassembled short reads were assembled with the help of SKESA (version 2.4.0). Quality control was carried out using Pathogenwatch (version 23.2.1) to filter out contaminated genomes belonging to the wrong species, resulting in 781 Enterobase downloaded genomes and 582

genomes from NCBI. The final dataset consisted of 1363 genomes which passed the quality control step and were collected in the years spanning from 1991-2024.

2.3 Feature Extraction

In order to develop a machine learning model well-equipped to predict gastric cancer status in *H. pylori* infected patients, a diverse set of features was included in the final dataset. These included clinical metadata from NCBI and Enterobase, gene presence and absence information obtained with the help of BLASTN-based homology searches, aggregated variant annotation features from VCF files and genome-derived sequence descriptors.

2.4 Clinical and Epidemiological features:

The clinical metadata associated with each *H.pylori* genome was collected from Enterobase and NCBI databases (Supplementary Table 1). The age, sex, and geographical location of the host along with their disease phenotype was extracted. With the help of clinical expertise, this disease phenotype was converted into one of two labels: gastric cancer and non-gastric cancer. This served as the label for the machine learning model and was appropriately encoded as mentioned in later sections.

2.5 Gene Presence and Absence features:

In addition to the clinical metadata, gene variant presence and absence information was also derived by identifying relevant gene variants associated with gastric and non-gastric cancer phenotype through exhaustive literature review.

These reference gene variant sequences linked to certain strains were extracted from a collection of genomes. These included cagA, babB, oipA, sabA and vacAs1m1 from *H.pylori* strain 60190;

babA2 from *H.pylori* strain ATCC 43504; hopQI from *H.pylori* strain J99; homB, hrgA, htrA, iceA1 from *H.pylori* strain 26695; hopQII from Tx30a.

Each of these assembled genomes were formatted into a separate BLAST compatible database. Each gene sequence was queried against each of the 1363 genome databases with the help of BLASTN (version 2.12.0). A gene was considered present in a genome if the top hit satisfied the below criteria:

- Percentage identity >= 90%
- Query coverage >= 80%
- E-value <= 1e-5

The presence of each gene was encoded as 1 and the absence as 0, resulting in a binary feature set (Supplementary Table 2). These thresholds were selected in order to balance sensitivity and specificity and to ensure accurate homologous gene detection by minimizing false positives.

2.6 Assembled Genomic Sequence Derived features:

To capture additional biological information, numerical sequence descriptors were extracted directly from the genomes using two Python-based packages, iFeatureOmega (version 1.0.2) and MathFeature (version 1.0) (Table 1).

These sequence-based features were computed by the packages at a contig level, which were then averaged in a length-weighted manner to generate a single representative feature vector per genome (Supplementary Table 3).

Specifically, for a given feature f, the genome-level feature average F was calculated as:

$$F = \frac{\sum_{i=1}^{n} f_i \times l_i}{\sum_{i=1}^{n} l_i} \quad (1)$$

where:

- f_i is the feature value for the i-the contig,
- l_i is the length (in base pairs) of the i-th contig,
- and n is the total number of contigs for that genome.

This method ensured that longer contigs which typically provide more reliable sequence information would contribute proportionally to the final feature representation.

Table 1: Feature types utilised in this study and the respective package

Package	Feature type
iFeatureOmega	Z-curve (144-bit)
	Kmer Type 1
	RCKmer Type 1
	Nucleic Acid Composition (NAC)
	Multivariate Mutual Information (MMI)
	CKSNAP type 1
MathFeature	Shannon Entropy

	ORF-Based Features

2.7 Variant-Based Features from VCF files:

Variant level features were extracted from genome-wide mutation profiles to include impact about sequence variation and their functional effects. The strain 26695 was used as a reference to call variants using Snippy (version 4.6.0), a rapid bacterial variant calling pipeline, on the assembled *H.pylori* genomes. Following variant calling, the Snippy-created VCF files were passed to another tool known as bcftools (version 1.2.1) for identifying the number of SNPs and MNPs. Using bcftools' (version 1.2.1) stats module, the raw counts of the following features were extracted SNPs, MNPs, Indels, and all possible nucleotide changes (e.g.,A → T, T → C, etc.). The variant effect functional annotations were also extracted from the VCF files to classify sequence changes based on their predicted effects on coding regions. The total counts of each variant effect were calculated per genome. These included types of mutations such as missense variants, frameshift mutations, mutations that affect start codons and combinations of these. For each genome, the number of variants falling into each functional category was counted and recorded as individual features.

These metrics were aggregated into per-genome features to help capture unexplored signals associated with the gastric cancer phenotype (Supplementary Table 4).

2.8 Data Preprocessing:

A comprehensive series of steps were carried out to preprocess the dataset to make it amenable for model training (Supplementary Table 5). The versions of scikit-learn, Pandas, matplotlib and Python used were 1.6.1, 2.2.3, 3.10.1 and 3.12.3.

Minor corrections were carried out to standardise categorical entries including correcting typographical errors, standardising the case of the characters and removing entries where the host sex was indeterminate, resulting in 1362 genomes.

The dataset was then split into training sets and test sets using the StratifiedShuffleSplit module from scikit-learn (version 1.6.1) (80% training set, 20% test set). For the input features, preprocessing for the input features involved normalisation using the MinMaxScaler module and one-hot encoding using the OneHotEncoder module in scikit-learn.

Label encoding was used to convert the categorical class labels of 'Gastric cancer' and 'Non-gastric cancer' into numeric representations using the LabelEncoder module in scikit-learn. Since the LabelEncoder module encodes the classes in alphabetical order, 'Gastric cancer' was encoded as 0 and 'Non-gastric cancer' was encoded as 1.

The same processes of normalisation and encoding were also applied separately to the test set to ensure that the preprocessing steps were consistent across training and test datasets. The transformations performed on the test set were done independently to prevent any form of information leakage. The resulting dataset was rendered suitable for further feature selection and model training steps.

2.9 Feature Selection:

Feature selection was only performed for the black-box models (XGBoost and Random Forest) to mitigate overfitting, enhance generalization and improve the model's explainability. In contrast, the baseline model (Logistic Regression) was trained on the complete feature set to provide a consistent benchmark for comparison. The ShapRFECV module in the probatus package (version 3.1.2) which performed recursive feature elimination based on SHAP

importance scores was used to select features. The features with the lowest SHAP scores were iteratively removed to get the final feature set.

The optimal feature subset was selected by maximizing the base model's (LightGBMClassifier) performance used for performing feature selection. This resulted in a final set of 93 features, which was used to subset the original dataset.

2.10 Model Training and Development

A Logistic Regression classifier was trained as the baseline model and chosen for its white-box functionality and for its ease of interpretability. A random state of 26 was used to maintain reproducibility.

Class imbalance in the dataset was handled using SMOTE-NC. It was preferred over SMOTE due to its ability to handle both categorical and continuous features by generating synthetic samples for the minority class while retaining the categorical nature of categorical variables. It was applied within a generated pipeline to the training folds within each cross-validation split to ensure that synthetic samples did not influence the model evaluation step on the validation step.

The full model training pipeline was built using imblearn's make_pipeline function and had SMOTE-NC and Logistic Regression classifier (with random state=26 to maintain reproducibility) as its two components.

Cross-validation was carried out using StratifiedShuffleSplit by splitting the dataset into 10 folds for robust evaluation, with the stratification step preserving the class distribution during both model training and evaluation. Recall was selected as the scoring metric to prioritise accurate identification of gastric cancer cases and minimise any false negatives. After cross-validation, the model was refitted on the entire training dataset to optimise its performance before applying it to the test set.

An XGBoost classifier was trained as the black-box model and chosen for its black-box high performance. The steps taken for preprocessing the dataset were identical to the baseline model, including the chosen random state.

Categorical feature indices were identified and used with SMOTE-NC and a pipeline to first oversample and then train an XGBClassifier on the reduced feature set. Hyperparameter tuning was carried out using the BayesSearchCV function from scikit-optimize package, which employs Bayesian optimization to automatically search for the best hyperparameters to enhance the performance of the model. Probability calibration was also performed on the XGBoost classifier specifically using sigmoid regression to ensure that the predicted probabilities better reflected the true likelihood of an outcome. After cross-validation, the model was refitted on the entire training dataset to optimise its performance before applying it to the test set.

A Random Forest black-box classifier was trained and evaluated in the same manner as the XGBoost model.

All the trained models were evaluated on the encoded test set using the trained pipeline to assess its generalisation abilities on unseen data.

2.11 Model Evaluation:

A classification report was generated in order to provide the values of key metrics like precision, recall and F1-score. Along with this, the Receiver Operating Characteristic (ROC) curve was plotted to detect a trade-off between correct identification of true positives and minimization of false positives. The Area Under the Curve (AUC) score was used to summarise the ROC curve, with an AUC value closer to 1 indicating better model performance. The Area Under the Precision-Recall curve (AUPRC) was also plotted to visualise the proportion of true positives among all predicted positives against the proportion of actual positives identified by the model.

The positive class was specified as gastric cancer, with the negative class being non-gastric cancer. Average precision (AP) score, computed as the area under the PR curve, offers a more effective evaluation than the ROC curve due to its sensitivity to class imbalance- the dominance of the non-gastric cancer class. A higher AP score is indicative of better performance in correctly classifying the minority class.

2.12 Post-hoc Explanation:

To interpret model predictions and understand the contribution of each feature, post-hoc explainability techniques were performed using SHAP Python package (version 0.47.1). The average impact of each feature on the model's output was visualized using SHAP summary plots for the three models (Logistic Regression (baseline), XGBoost, and Random Forest). These plots were used to illustrate the global feature importance across the whole dataset. These explainability techniques aid interpretation of models and help in assessing the consistency of feature contributions across different models.

3. Results:

Classification models were developed to group subjects into the gastric cancer and non-gastric cancer classes.

Logistic Regression was used as a baseline classifier to evaluate initial predictive performance on the binary classification task of gastric cancer vs non-gastric cancer. On applying oversampling, the baseline model achieved an overall accuracy of 0.81 on the test set. The recall for the gastric cancer class was 0.73 (Figure 2A), with a greater recall of 0.84 for the non-gastric

cancer class (Figure 2B). The weighted average recall of the model was moderately good at 0.81, and needs further refinement as recall is the most important factor in cancer prediction models where false negatives need to be minimized. The model also achieved a good weighted average precision of 0.82, with an AUPRC of 0.83 for gastric cancer and 0.94 for non-gastric cancer classes. The discriminatory power of the model denoted by an AUROC of 0.89 was equal for both the classes. These findings are illustrated in Figure 3A. Overall, logistic regression is a reasonable baseline with good performance, and also shows the necessity for more complex models that enhance the detection of the minority gastric-cancer class.

Both the black-box models, XGBoost and Random Forest displayed comparable performance with an increased discriminatory power, demonstrating an AUROC of 0.95 for both the classes. They also had superior gastric cancer associated AUPRC scores of 0.91 and 0.92 for XGBoost and Random Forest respectively as illustrated in Figures 3B and 3C respectively. The weighted average recall of 0.89 after the calibration of output probabilities was equal for both XGBoost and Random Forest models, while XGBoost had a 1% increase to 0.83 in gastric cancer specific recall compared to Random Forest's gastric cancer recall of 0.82. Both classifiers demonstrated a good balance between precision and recall. For the Random Forest classifier, the gastric cancer class associated precision was 0.84 and recall was 0.82, while for XGBoost, both gastric cancer class associated precision and recall were 0.83 (Figure 2A). The precision and recall associated with the non-gastric cancer class were higher with XGBoost achieving scores of 0.93 for both the metrics while Random Forest achieved precision of 0.92 and recall of 0.93 (Figure 2B). These results indicate that the models were effective at finding more true positives while maintaining a relatively low rate of false positives.

DeLong's test was performed to statistically evaluate the differences between the areas under two ROC curves derived from distinct models. The difference in the AUROC of XGBoostClassifier and the baseline Logistic Regression was statistically significant with a p-value equal to 0.0017 and Z score equal to -3.13. Similarly, the difference in AUROC of Random Forest and the baseline Logistic Regression was statistically significant with a p-value of 0.0013 and Z score of -3.20.

4. Discussion:

The SHAP summary plots (Figure 4) provide post-hoc global explanations of the model to identify the features that contributed most to the model's predictions. It can be seen that age was consistently identified as the top predictor of gastric cancer across all three models, aligning with existing scientific literature implicating older patients as more susceptible to developing gastric cancer. Loss of stop codons in high numbers was also associated with non-gastric cancer outcomes like peptic ulcer and chronic gastritis across all the models. The top 3 predictors in the baseline model were age, being from Europe and silent mutations in the start codons (Figure 4A). While higher values of age were found to be associated with gastric cancer, higher number of silent mutations in start codons and European origin were linked to less severe infection outcomes.

The top 3 contributing factors in the XGBoost model were age, the reverse kmer CAC and the simultaneous presence of frameshift mutations and loss of stop codons (Figure 4B). Higher age continued to be associated with gastric cancer while higher counts of CAC and the combined variant caused the model to predict benign outcomes.

The top 3 contributing factors in the Random Forest model were age, the reverse kmer CAC and the presence of AT rich regions (Figure 4C). Higher age and the presence of AT-rich regions

were associated with gastric cancer while higher counts of CAC caused the model to predict benign outcomes.

The two black-box models also consistently identified the presence of the CAC and AAA reverse kmers as associated with increased risk of non-gastric cancer outcomes, while the ATA reverse kmer was associated with developing gastric cancer as the clinical outcome. Combinations of variants have also been marked as gastric-cancer inducing factors in both the XGBoost and Random Forest models, such as the high presence of all 4 mutations- frameshift variants, loss of stop codons, missense variants and splice region variants.

Although the models possessed remarkable discriminatory power, a variety of pertinent features were not incorporated into the model. It has been shown that family history, smoking habits, diet and other clinical features play a very important role in the development of *H. pylori* infection associated outcomes, along with the role of host immune factors[30]. The data also had a large number of European samples (870), and a wider geographic distribution of publicly available sample data could render the model more generalizable.

5. Conclusions

The two black-box models achieved greater recall in the prediction of gastric cancer compared to the baseline logistic regression model. The top predictor was consistently found to be age across all three models, while different genomic markers were identified by each of the models. Incorporating genomic features of the host along with clinical metadata could enhance the predictive power and the robustness of the models since current *H. pylori* prediction models focus largely on clinical features . The integration of transcriptomic data and host gut metagenomic data could account for gene expression and interaction of *H. pylori* with other gut

microbes. The sequence-derived genomic features implicated in this study open avenues for further research into the *H. pylori* genome beyond the already implicated virulence genes to assess possible causation behind the associations found in this study.

6. List of abbreviations

H. pylori: *Helicobacter pylori*

XGBoost: eXtreme Gradient Boosting

SMOTE-NC: Synthetic Minority Over-sampling Technique for Nominal and Continuous

SHAP: SHapley Additive exPlanations

AUROC: Area Under Receiver Operating Characteristic curve

MALT: Mucosa-associated lymphoid tissue

WHO: World Health Organisation

cagA: cytotoxin-associated gene A

cagPAI: cag pathogenicity island

vacA: vacuolating cytotoxin A

BabA: blood group antigen binding adhesin

oipA: outer inflammatory protein A

sLe^X : sialylated Lewis antigens

SRA: Sequence Read Archive

NAC: Nucleic Acid Composition

MMI: Multivariate Mutual Information

ROC: Receiver Operating Characteristic

AUC: Area Under the Curve

AUPRC: Area Under the Precision-Recall curve

AP: Average precision

7. Declarations

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable

Availability of data and materials:

The *Helicobacter pylori* genomes analyzed in this study are publicly available on the National Center for Biotechnology Information (NCBI) GenBank repository (https://www.ncbi.nlm.nih.gov/genbank/) and EnteroBase (https://enterobase.warwick.ac.uk/). Accession numbers for all 1,363 genomes used in analyses are provided in Supplementary Table 1. The custom code and analysis pipelines used in this study are available on GitHub (https://github.com/Venkatesh-99/HP_ML/) and archived in Zenodo at (https://doi.org/10.5281/zenodo.16598966).

Competing interests:

The authors declare that they have no competing interests.

Funding:

No funding source was utilized in carrying out this study.

Authors' contributions:

SPW and VN contributed equally to conceptualization, data curation, investigation and methodology design, software and formal analysis, data visualization, validation, writing the original draft, reviewing and editing the manuscript. MPT contributed to data curation and formal analysis. JJJ, NV and GGT contributed to conceptualization, reviewing and editing the manuscript. BV contributed to reviewing and editing the manuscript, provision of resources, project administration and supervision. All authors read and approved the final manuscript.

Acknowledgements:

Not applicable

References:

1. Shatila, M. & Thomas, A. S. Current and Future Perspectives in the Diagnosis and Management of Helicobacter pylori Infection. *J. Clin. Med.* **11**, 5086 (2022).
2. Elbehiry, A. *et al.* Helicobacter pylori Infection: Current Status and Future Prospects on Diagnostic, Therapeutic and Control Challenges. *Antibiot. Basel Switz.* **12**, 191 (2023).
3. Pucułek, M. *et al.* Helicobacter pylori associated factors in the development of gastric cancer with special reference to the early-onset subtype. *Oncotarget* **9**, 31146–31162 (2018).
4. Backert, S. & Blaser, M. J. The Role of CagA in the Gastric Biology of Helicobacter pylori. *Cancer Res.* **76**, 4028–4031 (2016).
5. Kao, C.-Y., Sheu, B.-S. & Wu, J.-J. Helicobacter pylori infection: An overview of bacterial virulence factors and pathogenesis. *Biomed. J.* **39**, 14–23 (2016).
6. McClain, M. S., Beckett, A. C. & Cover, T. L. Helicobacter pylori Vacuolating Toxin and Gastric Cancer. *Toxins* **9**, 316 (2017).
7. Foegeding, N. J., Caston, R. R., McClain, M. S., Ohi, M. D. & Cover, T. L. An Overview of Helicobacter pylori VacA Toxin Biology. *Toxins* **8**, 173 (2016).
8. González, C. A. *et al.* Helicobacter pylori cagA and vacA genotypes as predictors of progression of gastric preneoplastic lesions: a long-term follow-up in a high-risk area in Spain. *Am. J. Gastroenterol.* **106**, 867–874 (2011).
9. Fujimoto, S. *et al.* Helicobacter pylori BabA expression, gastric mucosal injury, and clinical outcome. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **5**, 49–58 (2007).
10. Gerhard, M. *et al.* Clinical relevance of the Helicobacter pylori gene for blood-group antigen-binding adhesin. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 12778–12783 (1999).

11. Chiarini, A. *et al.* Prevalence of virulence-associated genotypes of Helicobacter pylori and correlation with severity of gastric pathology in patients from western Sicily, Italy. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **28**, 437–446 (2009).
12. Teymournejad, O., Mobarez, A. M., Hassan, Z. M. & Talebi Bezmin Abadi, A. Binding of the Helicobacter pylori OipA causes apoptosis of host cells via modulation of Bax/Bcl-2 levels. *Sci. Rep.* **7**, 8036 (2017).
13. Dossumbekova, A. *et al.* Helicobacter pylori HopH (OipA) and bacterial pathogenicity: genetic and functional genomic analysis of hopH gene polymorphisms. *J. Infect. Dis.* **194**, 1346–1355 (2006).
14. Zambon, C.-F., Navaglia, F., Basso, D., Rugge, M. & Plebani, M. Helicobacter pylori babA2, cagA, and s1 vacA genes work synergistically in causing intestinal metaplasia. *J. Clin. Pathol.* **56**, 287–291 (2003).
15. Markovska, R., Boyanova, L., Yordanov, D., Gergova, G. & Mitov, I. Helicobacter pylori oipA genetic diversity and its associations with both disease and cagA, vacA s, m, and i alleles among Bulgarian patients. *Diagn. Microbiol. Infect. Dis.* **71**, 335–340 (2011).
16. Ando, T. *et al.* Polymorphisms of Helicobacter pylori HP0638 reflect geographic origin and correlate with cagA status. *J. Clin. Microbiol.* **40**, 239–246 (2002).
17. Oleastro, M. *et al.* Evaluation of the clinical significance of homB, a novel candidate marker of Helicobacter pylori strains associated with peptic ulcer disease. *J. Infect. Dis.* **198**, 1379–1387 (2008).
18. Hage, N. *et al.* Structural basis of Lewis(b) antigen binding by the Helicobacter pylori adhesin BabA. *Sci. Adv.* **1**, e1500315 (2015).
19. Doohan, D., Rezkitha, Y. A. A., Waskito, L. A., Yamaoka, Y. & Miftahussurur, M.

Helicobacter pylori BabA-SabA Key Roles in the Adherence Phase: The Synergic Mechanism for Successful Colonization and Disease Development. *Toxins* **13**, 485 (2021).

20. Unemo, M. *et al.* The sialic acid binding SabA adhesin of Helicobacter pylori is essential for nonopsonic activation of human neutrophils. *J. Biol. Chem.* **280**, 15390–15397 (2005).
21. Perna, A. M. *et al.* Fragment-Based De Novo Design Reveals a Small-Molecule Inhibitor of Helicobacter Pylori HtrA. *Angew. Chem. Int. Ed Engl.* **54**, 10244–10248 (2015).
22. Hoy, B. *et al.* Helicobacter pylori HtrA is a new secreted virulence factor that cleaves E-cadherin to disrupt intercellular adhesion. *EMBO Rep.* **11**, 798–804 (2010).
23. Dadashzadeh, K., Peppelenbosch, M. P. & Adamu, A. I. Helicobacter pylori Pathogenicity Factors Related to Gastric Cancer. *Can. J. Gastroenterol. Hepatol.* **2017**, 7942489 (2017).
24. Leylabadlo, H. E., Yekani, M. & Ghotaslou, R. Helicobacter pylori hopQ alleles (type I and II) in gastric cancer. *Biomed. Rep.* **4**, 601–604 (2016).
25. Malfertheiner, P. *et al.* Helicobacter pylori infection. *Nat. Rev. Dis. Primer* **9**, 19 (2023).
26. Alm, R. A. & Noonan, B. The Genome. in *Helicobacter pylori: Physiology and Genetics* (eds. Mobley, H. L., Mendz, G. L. & Hazell, S. L.) (ASM Press, Washington (DC), 2001).
27. Taninaga, J. *et al.* Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. *Sci. Rep.* **9**, 12384 (2019).
28. Jiang, S. *et al.* Machine learning: A non-invasive prediction method for gastric cancer based on a survey of lifestyle behaviors. *Front. Artif. Intell.* **5**, (2022).
29. Afrash, M. R., Shafiee, M. & Kazemi-Arpanahi, H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. *BMC Gastroenterol.* **23**, 6 (2023).
30. Öztekin, M., Yılmaz, B., Ağagündüz, D. & Capasso, R. Overview of Helicobacter pylori

Infection: Clinical Features, Treatment, and Nutritional Aspects. *Dis. Basel Switz.* **9**, 66 (2021).

Additional Files:

- **Supplementary_Table_1.xlsx**: Clinical metadata for each sample, including host age, sex, geographic location, and labeled disease outcome (gastric vs non-gastric cancer).
- **Supplementary_Table_2.xlsx**: Binary matrix indicating presence (1) or absence (0) of known H. pylori virulence genes per genome, based on BLASTN results.
- **Supplementary_Table_3.xlsx**: Length-weighted numerical genome descriptors computed using iFeatureOmega and MathFeature for each H. pylori genome.
- **Supplementary_Table_4.xlsx**: Aggregated variant-level features including mutation types and functional annotations extracted from VCF files for each genome.
- **Supplementary_Table_5.xlsx**: Final merged dataset containing all the extracted features before model training.

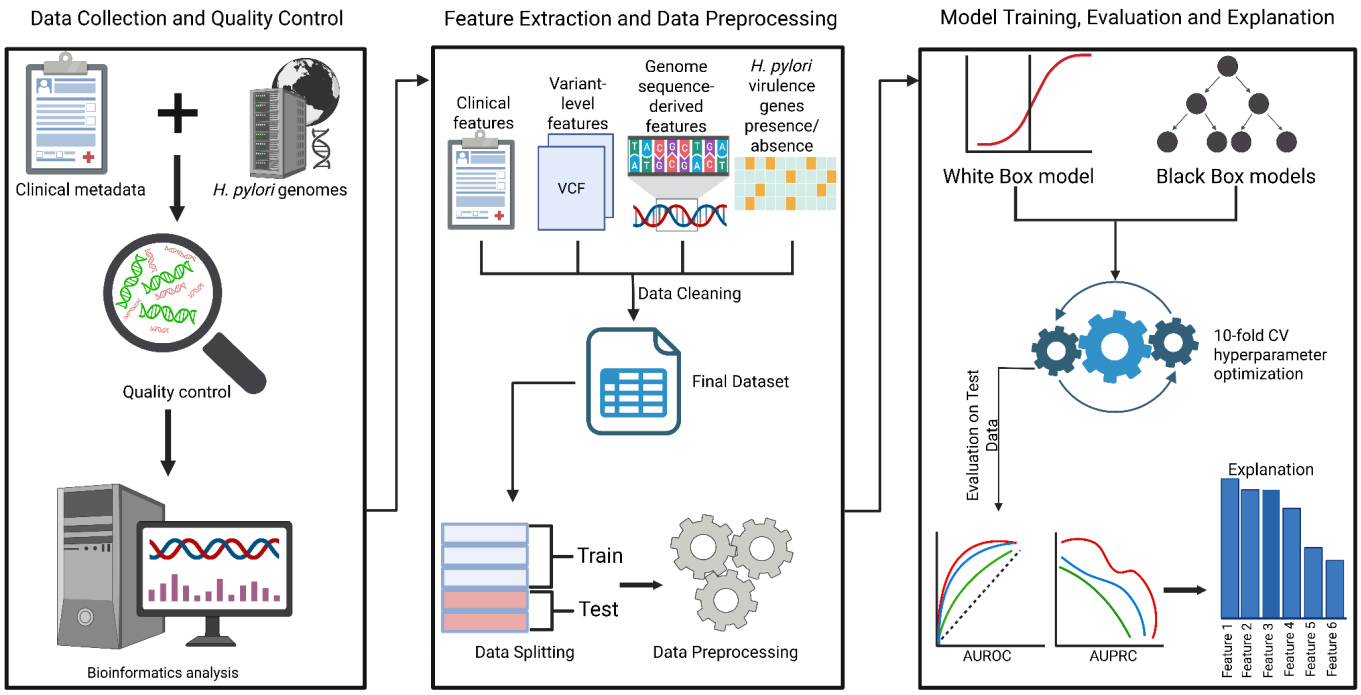


Figure 1: Overview of the machine learning workflow for classification of gastric cancer and non-gastric cancer using *H. pylori* genomes and clinical data. The workflow consists of three stages, depicted from left to right, with arrows indicating the flow of the stages: (1) Data collection and quality control, host clinical metadata and *H. pylori* whole genomes are integrated and filtered for quality, and bioinformatics analysis (variant calling); (2) Feature extraction and data preprocessing (normalization, one-hot encoding, label encoding), including clinical features, variant-level features from VCF files, and genome sequence-derived features and presence/absence data of *H. pylori* virulence genes to construct the final dataset, followed by data preprocessing; and (3) Model training, evaluation and explanation, involving binary classification using white-box (Logistic Regression) and black-box models (XGBoost and Random Forest) with 10-fold cross-validation and hyperparameter bayesian optimization, performance evaluation using AUROC and AUPRC metrics, and finally feature importance interpretation.

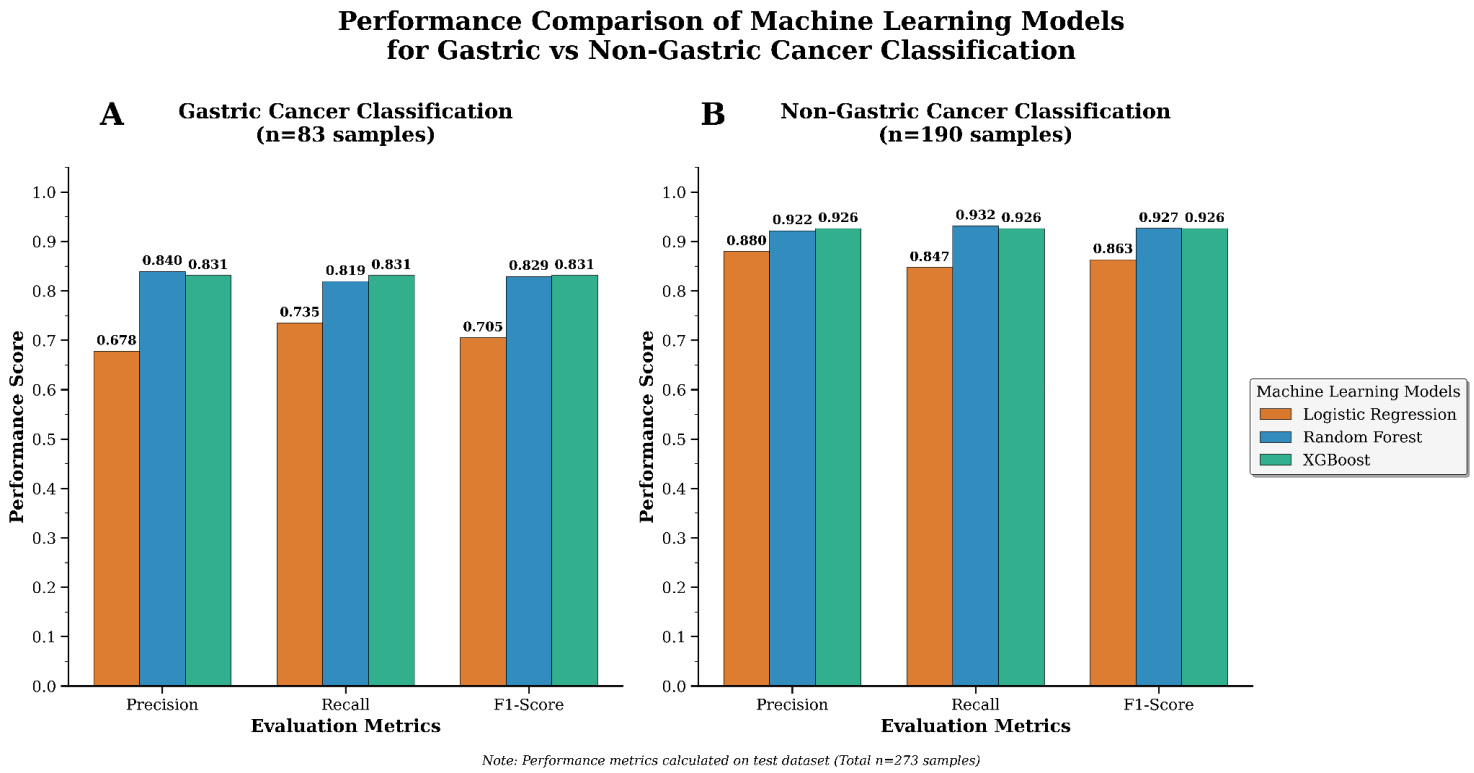


Figure 2: Performance comparison of machine learning models for gastric versus non-gastric cancer classification. Bar charts show precision, recall, and F1-score metrics for three algorithms: Logistic Regression (orange), Random Forest (blue), and XGBoost (green). Panel A presents results for gastric cancer classification (n=83 samples) and Panel B shows non-gastric cancer classification (n=190 samples). All models demonstrate high performance across metrics (>0.8), with Random Forest and XGBoost showing marginally superior performance compared to Logistic Regression. Performance metrics were calculated using standard evaluation protocols.

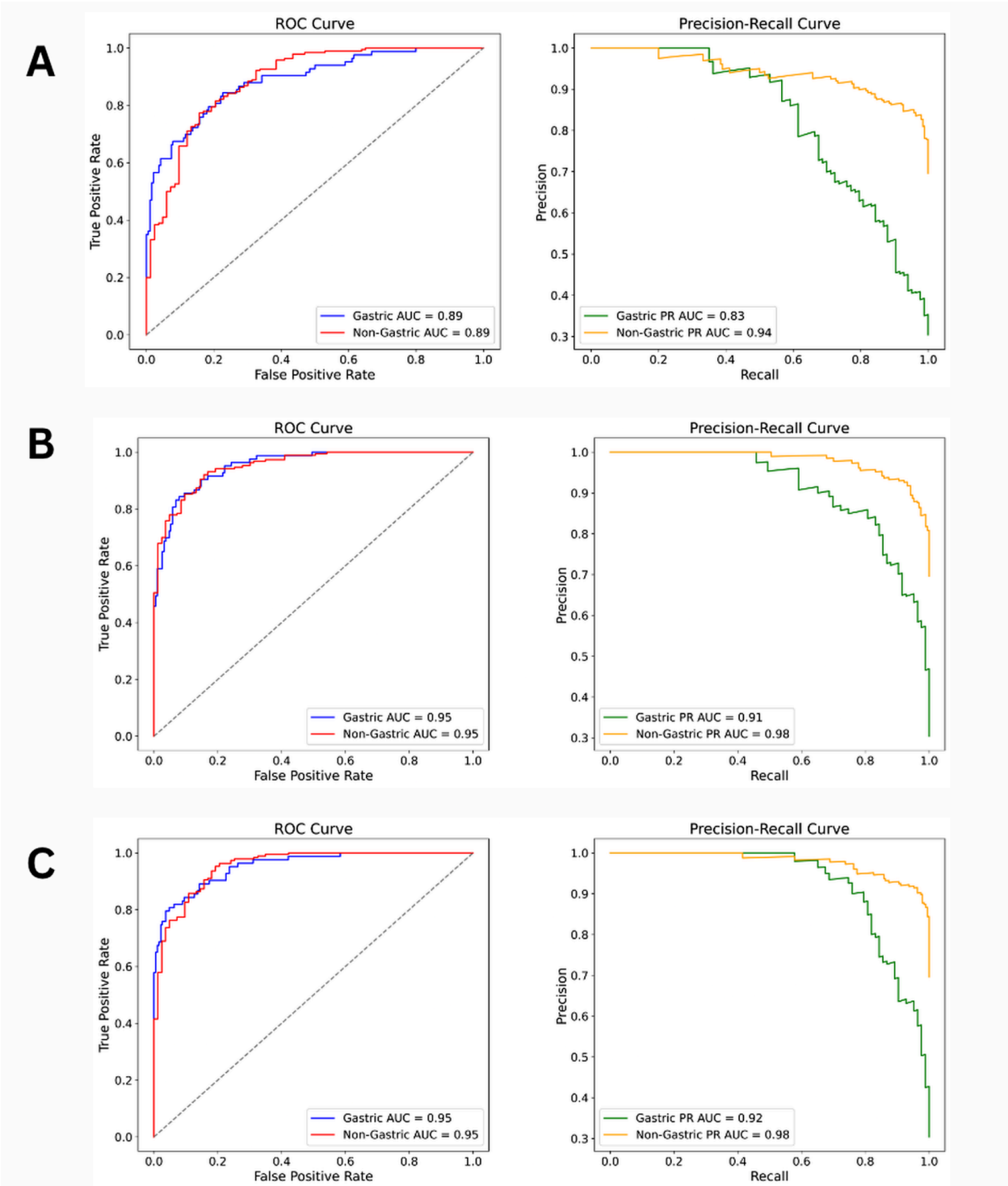


Figure 3: **Receiver operating characteristic (ROC) and precision-recall (PR) curves for the classification of gastric cancer and non-gastric cancer samples.** (A) Logistic regression performance: AUROC = 0.89 for both classes; AUPRC = 0.83 (gastric) and 0.94 (non-gastric).

(B) XGBoostClassifier: AUROC = 0.95 for both classes; AUPRC = 0.91 (gastric) and 0.98 (non-gastric). (C) RandomForestClassifier: AUROC = 0.95 for both classes; AUPRC = 0.92 (gastric) and 0.98 (non-gastric).

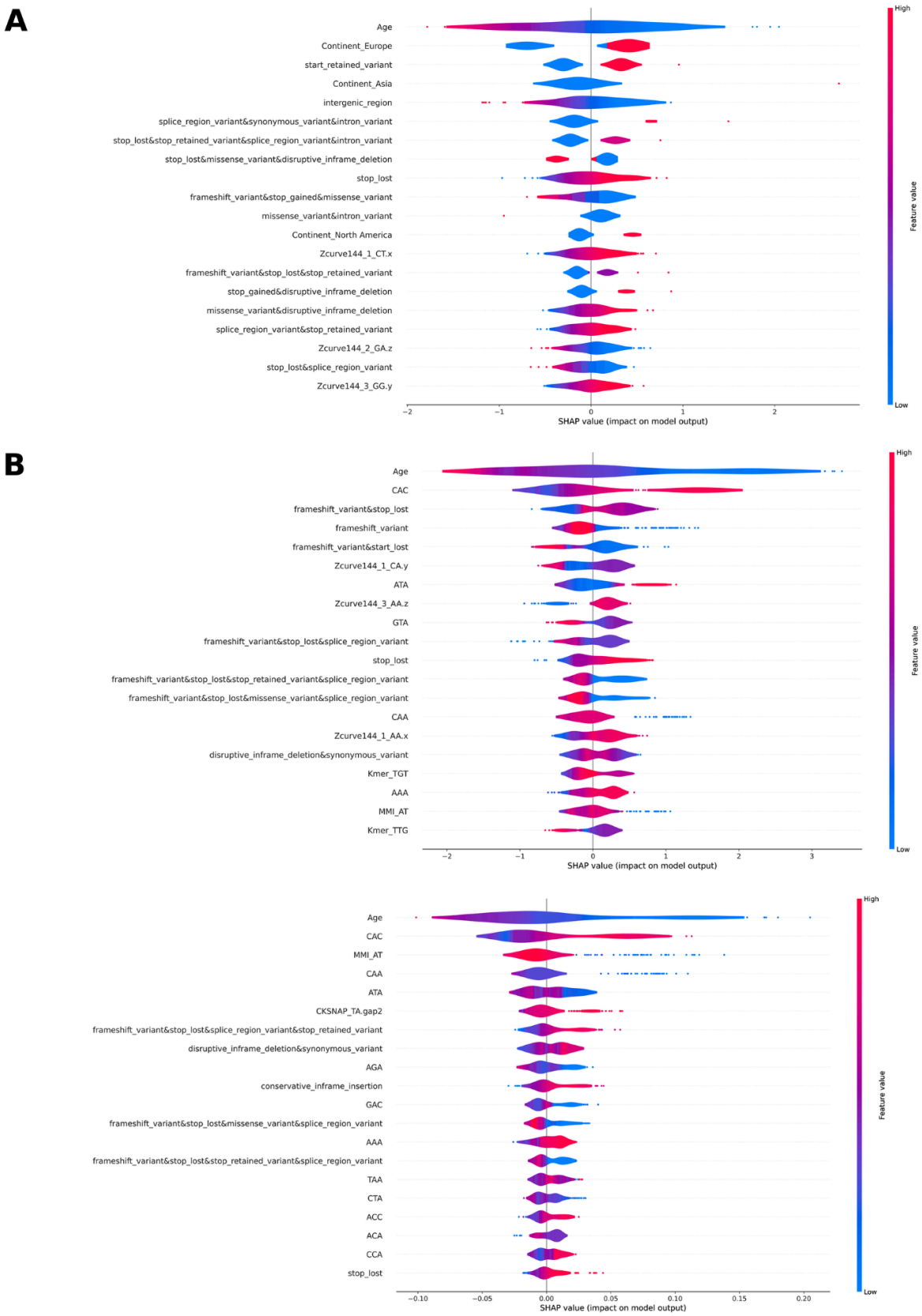
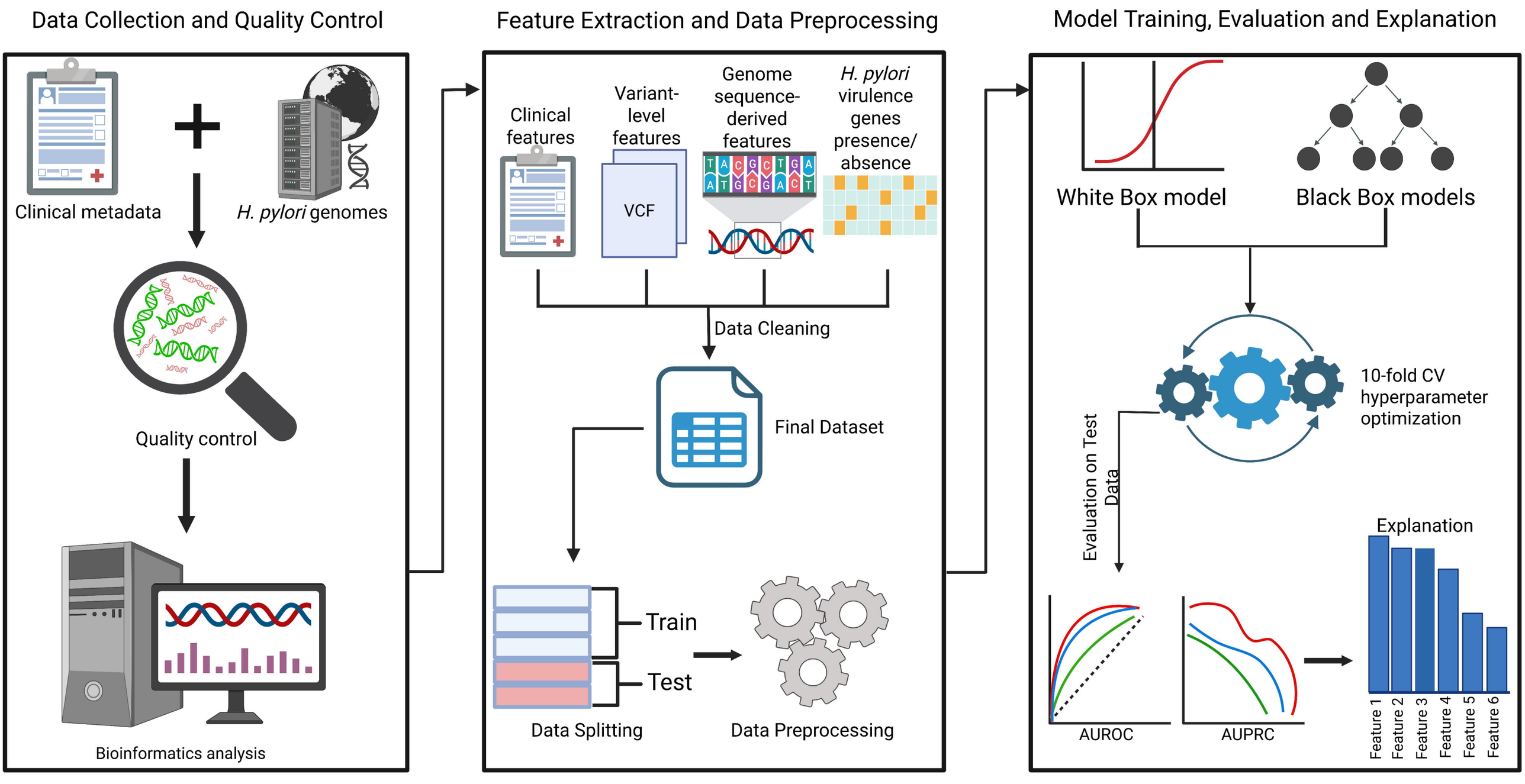
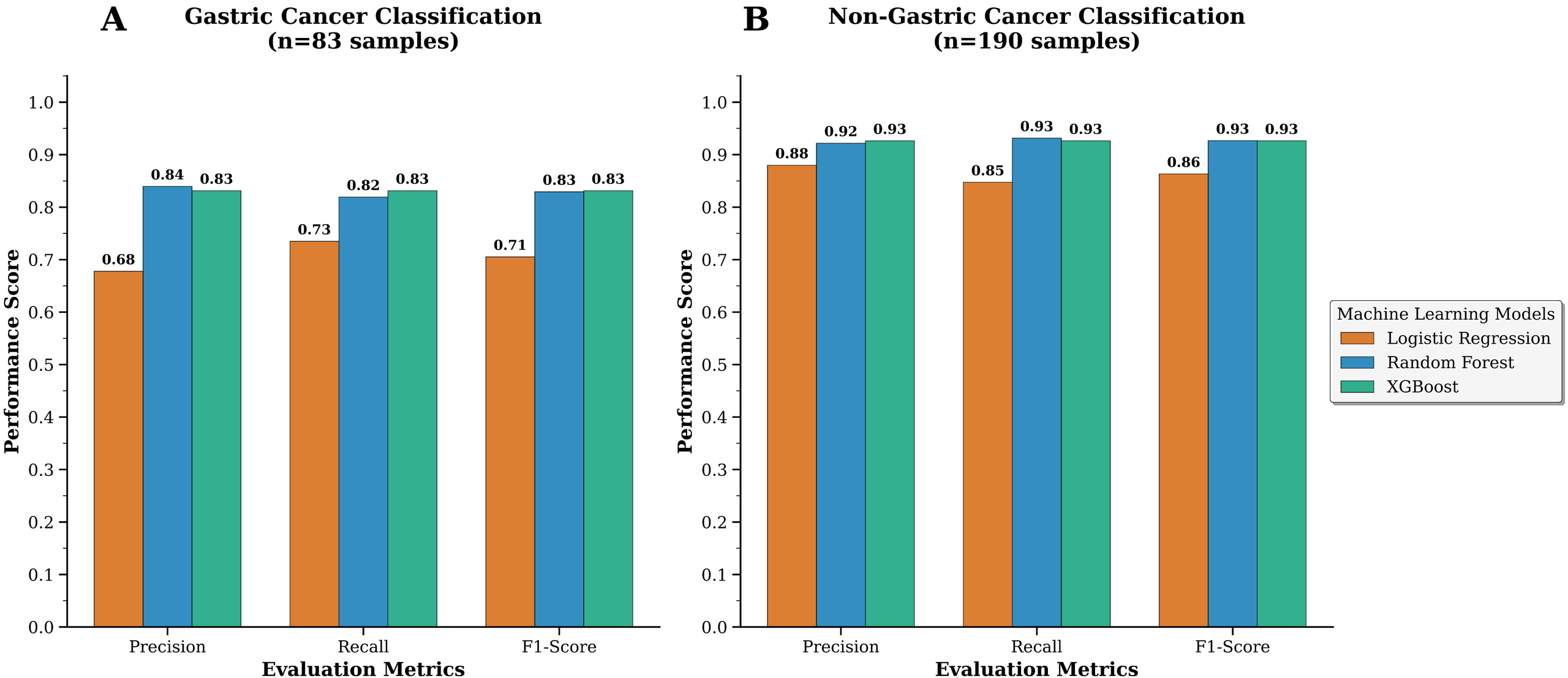

A
Age
Continent_Europe
start_retained_variant
Continent_Asia
intergenic_region
splice_region_variant&synonymous_variant&intron_variant
stop_lost&stop_retained_variant&splice_region_variant&intron_variant
stop_lost&missense_variant&disruptive_inframe_deletion
stop_lost
frameshift_variant&stop_gained&missense_variant
missense_variant&intron_variant
Continent_North America
Zcurve144_1_CT.x
frameshift_variant&stop_lost&stop_retained_variant
stop_gained&disruptive_inframe_deletion
missense_variant&disruptive_inframe_deletion
splice_region_variant&stop_retained_variant
Zcurve144_2_GA.z
stop_lost&splice_region_variant
Zcurve144_3_GG.y
−2 −1 0 1 2
SHAP value (impact on model output)
High
Feature value
Low
B
Age
CAC
frameshift_variant&stop_lost
frameshift_variant
frameshift_variant&start_lost
Zcurve144_1_CA.y
ATA
Zcurve144_3_AA.z
GTA
frameshift_variant&stop_lost&splice_region_variant
stop_lost
frameshift_variant&stop_lost&stop_retained_variant&splice_region_variant
frameshift_variant&stop_lost&missense_variant&splice_region_variant
CAA
Zcurve144_1_AA.x
disruptive_inframe_deletion&synonymous_variant
Kmer_TGT
AAA
MMI_AT
Kmer_TTG
−2 −1 0 1 2 3
SHAP value (impact on model output)
High
Feature value
Low
C
Age
CAC
MMI_AT
CAA
ATA
CKSNAP_TA.gap2
frameshift_variant&stop_lost&splice_region_variant&stop_retained_variant
disruptive_inframe_deletion&synonymous_variant
AGA
conservative_inframe_insertion
GAC
frameshift_variant&stop_lost&missense_variant&splice_region_variant
AAA
frameshift_variant&stop_lost&stop_retained_variant&splice_region_variant
TAA
CTA
ACC
ACA
CCA
stop_lost
−0.10 −0.05 0.00 0.05 0.10 0.15 0.20
SHAP value (impact on model output)
High
Feature value
Low

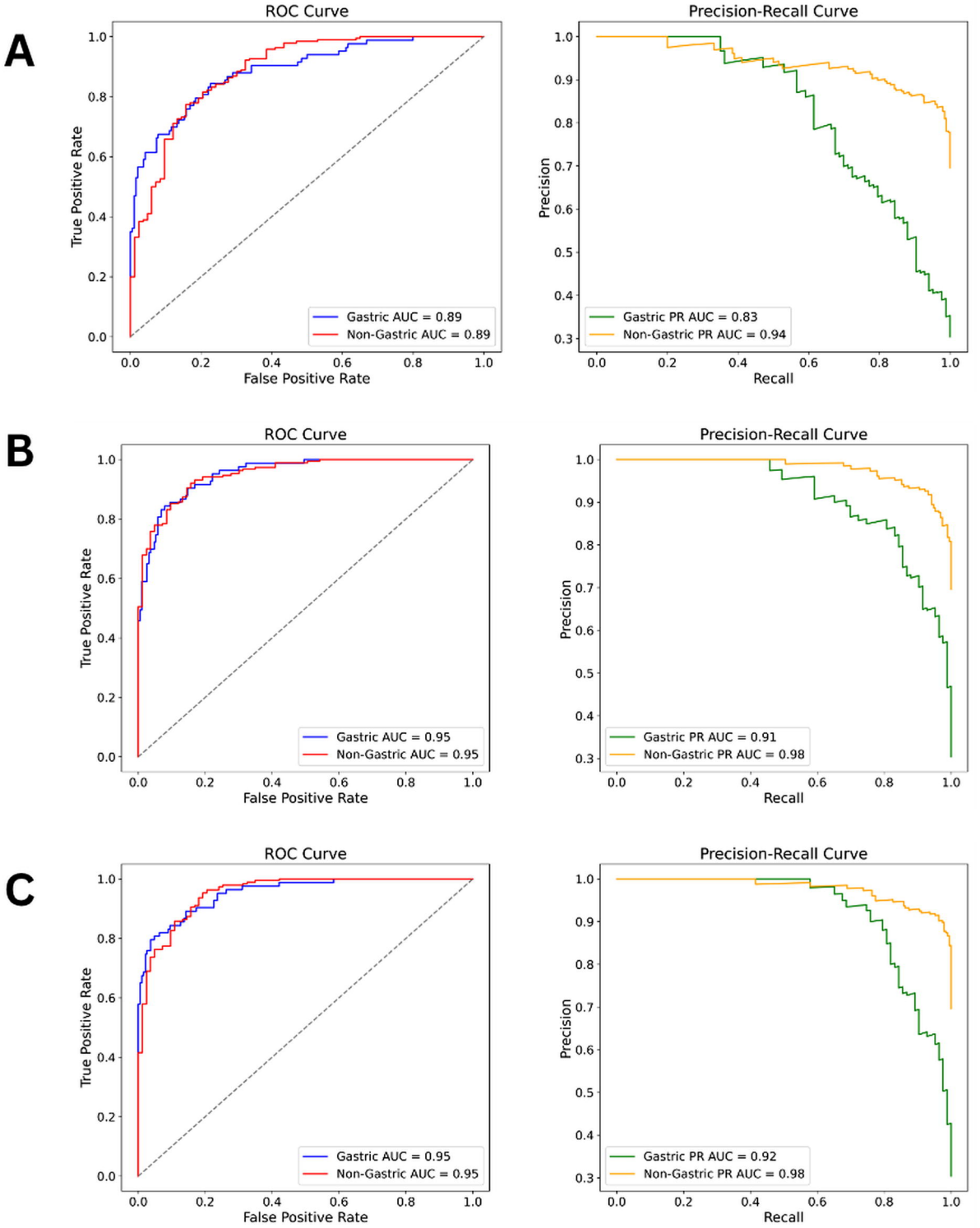
Figure 4: SHAP summary plots showing global feature importance for all three predictive models. (A) Logistic Regression, (B) XGBoost, and (C) Random Forest. The SHAP value on the x-axis represents the feature's impact on the model's output, with violin plots illustrating the distribution of SHAP values across all samples for each feature. The colour gradient indicates the original value of the feature (blue = low, red = high). Features are ranked by their mean absolute SHAP value, from top (most important) to bottom. These plots highlight which features most strongly influence the model predictions.

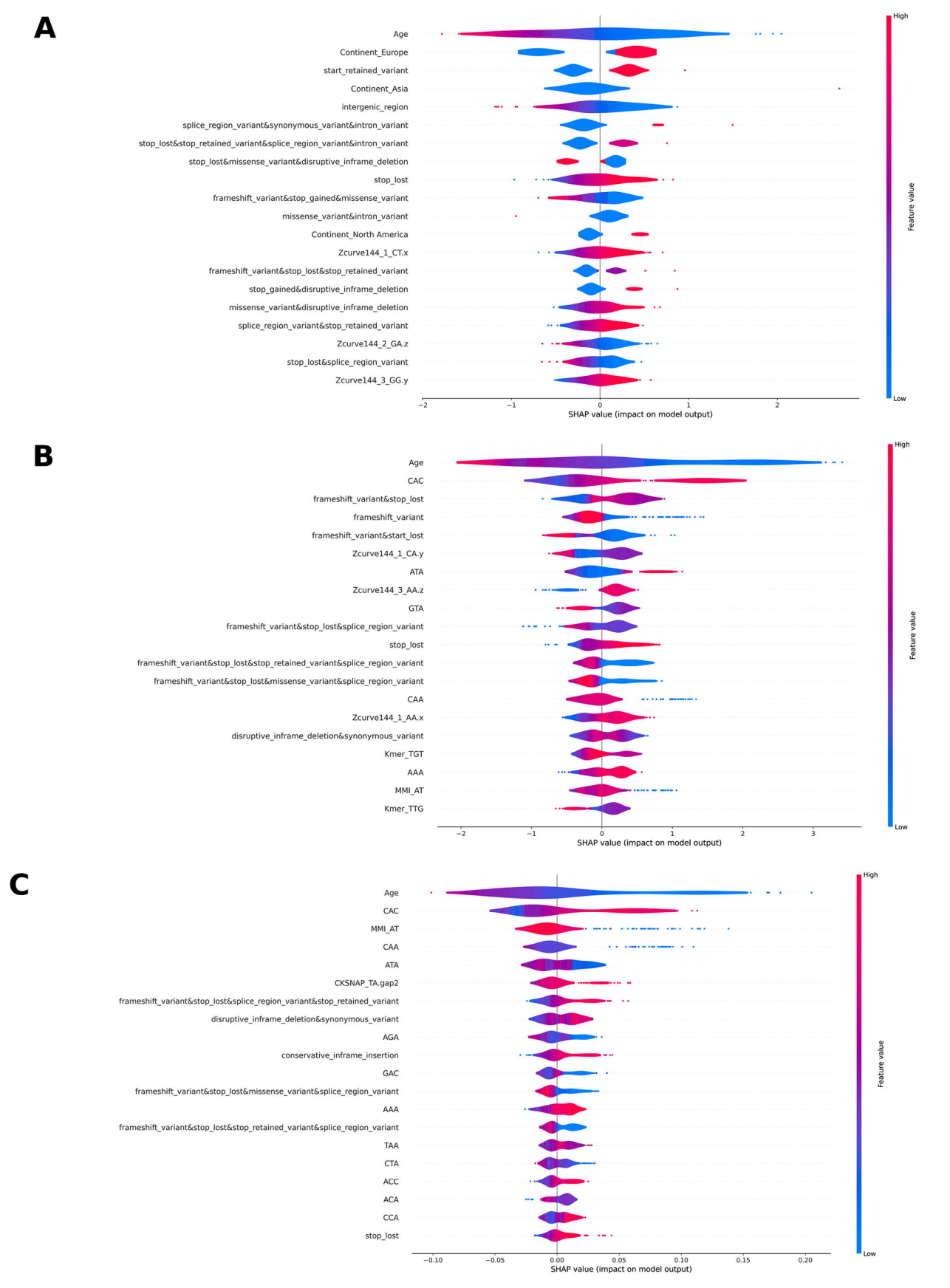

Data Collection and Quality Control
Clinical metadata
H. pylori genomes
Quality control
Bioinformatics analysis
Feature Extraction and Data Preprocessing
Clinical features
Variant-level features
VCF
Genome sequence-derived features
H. pylori virulence genes presence/ absence
Data Cleaning
Final Dataset
Train
Test
Data Splitting
Data Preprocessing
Model Training, Evaluation and Explanation
White Box model
Black Box models
10-fold CV hyperparameter optimization
Evaluation on Test Data
AUROC
AUPRC
Explanation
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Feature 6

Performance Comparison of Machine Learning Models for Gastric vs Non-Gastric Cancer Classification



Note: Performance metrics calculated on test dataset (Total n=273 samples)


A
ROC Curve
True Positive Rate
False Positive Rate
Gastric AUC = 0.89
Non-Gastric AUC = 0.89
Precision-Recall Curve
Precision
Recall
Gastric PR AUC = 0.83
Non-Gastric PR AUC = 0.94
B
ROC Curve
True Positive Rate
False Positive Rate
Gastric AUC = 0.95
Non-Gastric AUC = 0.95
Precision-Recall Curve
Precision
Recall
Gastric PR AUC = 0.91
Non-Gastric PR AUC = 0.98
C
ROC Curve
True Positive Rate
False Positive Rate
Gastric AUC = 0.95
Non-Gastric AUC = 0.95
Precision-Recall Curve
Precision
Recall
Gastric PR AUC = 0.92
Non-Gastric PR AUC = 0.98


A
Age
Continent_Europe
start_retained_variant
Continent_Asia
intergenic_region
splice_region_variant&synonymous_variant&intron_variant
stop_lost&stop_retained_variant&splice_region_variant&intron_variant
stop_lost&missense_variant&disruptive_inframe_deletion
stop_lost
frameshift_variant&stop_gained&missense_variant
missense_variant&intron_variant
Continent_North America
Zcurve144_1_CT.x
frameshift_variant&stop_lost&stop_retained_variant
stop_gained&disruptive_inframe_deletion
missense_variant&disruptive_inframe_deletion
splice_region_variant&stop_retained_variant
Zcurve144_2_GA.z
stop_lost&splice_region_variant
Zcurve144_3_GG.y
−2
−1
0
1
2
SHAP value (impact on model output)
High
Feature value
Low
B
Age
CAC
frameshift_variant&stop_lost
frameshift_variant
frameshift_variant&start_lost
Zcurve144_1_CA.y
ATA
Zcurve144_3_AA.z
GTA
frameshift_variant&stop_lost&splice_region_variant
stop_lost
frameshift_variant&stop_lost&stop_retained_variant&splice_region_variant
frameshift_variant&stop_lost&missense_variant&splice_region_variant
CAA
Zcurve144_1_AA.x
disruptive_inframe_deletion&synonymous_variant
Kmer_TGT
AAA
MMI_AT
Kmer_TTG
−2
−1
0
1
2
3
SHAP value (impact on model output)
High
Feature value
Low
C
Age
CAC
MMI_AT
CAA
ATA
CKSNAP_TA.gap2
frameshift_variant&stop_lost&splice_region_variant&stop_retained_variant
disruptive_inframe_deletion&synonymous_variant
AGA
conservative_inframe_insertion
GAC
frameshift_variant&stop_lost&missense_variant&splice_region_variant
AAA
frameshift_variant&stop_lost&stop_retained_variant&splice_region_variant
TAA
CTA
ACC
ACA
CCA
stop_lost
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
SHAP value (impact on model output)
High
Feature value
Low