Predicting hosts and cross-species transmission of Streptococcus agalactiae by interpretable machine learning

Yunxiao Ren, Carmen Li, Dulmini Nanayakkara Sapugahawatte, Chendi Zhu, Sebastian Spänig, Dorota Jamrozy, Julian Rothen, Claudia A. Daubenberger, Stephen D. Bentley, Margaret Ip, Dominik Heider

Article - Version of Record

Computers in Biology and Medicine 171 (2024) 108185

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Predicting hosts and cross-species transmission of *Streptococcus agalactiae* by interpretable machine learning

Yunxiao Ren [a,1], Carmen Li [b,1], Dulmini Nanayakkara Sapugahawatte [b], Chendi Zhu [b], Sebastian Spänig [a], Dorota Jamrozy [c], Julian Rothen [d,e], Claudia A. Daubenberger [d,e], Stephen D. Bentley [c], Margaret Ip [b,**], Dominik Heider [a,f,g,*]

[a] Department for Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany
[b] Department of Microbiology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China
[c] Parasites and Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
[d] Swiss Tropical and Public Health Institute (Swiss TPH) Basel, Department of Medical Parasitology and Infection Biology, 4002, Basel, Switzerland
[e] University of Basel, 4002, Basel, Switzerland
[f] Institute for Computer Science, University of Düsseldorf, 40211, Düsseldorf, Germany
[g] Center for Digital Health, Heinrich Heine University Düsseldorf, Moorenstr. 5, 40225, Düsseldorf, Germany

ARTICLE INFO

Keywords:
Hosts prediction
Host adaptations
Cross-species transmission
Interpretable machine learning

ABSTRACT

Background: *Streptococcus agalactiae*, commonly known as Group B *Streptococcus* (GBS), exhibits a broad host range, manifesting as both a beneficial commensal and an opportunistic pathogen across various species. In humans, it poses significant risks, causing neonatal sepsis and meningitis, along with severe infections in adults. Additionally, it impacts livestock by inducing mastitis in bovines and contributing to epidemic mortality in fish populations. Despite its wide host spectrum, the mechanisms enabling GBS to adapt to specific hosts remain inadequately elucidated. Therefore, the development of a rapid and accurate method differentiates GBS strains associated with particular animal hosts based on genome-wide information holds immense potential. Such a tool would not only bolster the identification and containment efforts during GBS outbreaks but also deepen our comprehension of the bacteria's host adaptations spanning humans, livestock, and other natural animal reservoirs.

Methods and results: Here, we developed three machine learning models—random forest (RF), logistic regression (LR), and support vector machine (SVM) based on genome-wide mutation data. These models enabled precise prediction of the host origin of GBS, accurately distinguishing between human, bovine, fish, and pig hosts. Moreover, we conducted an interpretable machine learning using SHapley Additive exPlanations (SHAP) and variant annotation to uncover the most influential genomic features and associated genes for each host. Additionally, by meticulously examining misclassified samples, we gained valuable insights into the dynamics of host transmission and the potential for zoonotic infections.

Conclusions: Our study underscores the effectiveness of random forest (RF) and logistic regression (LR) models based on mutation data for accurately predicting GBS host origins. Additionally, we identify the key features associated with each GBS host, thereby enhancing our understanding of the bacteria's host-specific adaptations.

1. Introduction

Streptococcus agalactiae, commonly referred to as Group B *Streptococcus* (GBS), typically resides in the gastrointestinal and genital tracts of healthy individuals. Although often benign, GBS can pose significant health risks, particularly to susceptible groups such as newborns, pregnant women, and the elderly, occasionally leading to severe infections [1–3]. GBS is a leading cause of neonatal infections, which can result in

* Corresponding author. Department of Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany.
** Corresponding author.
E-mail address: dominik.heider@hhu.de (D. Heider).
[1] These authors contributed equally: Yunxiao Ren, Carmen Li.

https://doi.org/10.1016/j.compbiomed.2024.108185
Received 2 November 2023; Received in revised form 13 February 2024; Accepted 18 February 2024
Available online 20 February 2024

severe outcomes such as sepsis, meningitis, and pneumonia. Newborns who contract GBS during delivery are at exceptionally high risk of developing these infections, which can be life-threatening [4]. To prevent GBS disease in newborns, many countries have implemented screening protocols for pregnant women and providing antibiotics during labor to those identified as carriers [5]. Additionally, GBS infections in pregnant can lead to chorioamnionitis, characterized by inflammation of the fetal membrane, potentially triggering preterm delivery and other associated complications [6]. Therefore, timely identification and management of GBS infection in pregnant women are crucial to prevent adverse outcomes for both the mother and the fetus. GBS is also a significant concern for older adults with weakened immune systems. These individuals are at increased risk of developing invasive GBS infections, such as osteomyelitis, arthritis, endocarditis, sepsis, and meningitis [7]. Furthermore, increased incidence of invasive GBS in healthy adults has also been observed, and mortality rate was associated with older adults (i.e., ≥65 years), with a 50% mortality rate in the elderly upon infection [7].

Moreover, GBS is also prevalent among animals, including livestock, pets, and wild animals [8–10]. While most GBS infections in humans are caused by person-to-person transmission, there is evidence that zoonotic transmission may also occur [11]. While the risk of zoonotic transmission of GBS appears to be relatively low, it is still a concern, particularly in high-risk populations. Implementing a host prediction model can help discern the primary host source of GBS, facilitating (i) the tracing of potential zoonotic origins, (ii) preventing the spread of GBS infections in both animal and human populations, and (iii) minimizing the impact on the public health of this adaptable bacterium.

Machine learning (ML) plays a pivotal role in various facets of biomedicine, encompassing tasks like predicting specific biological functional sequences [12–14], diagnosing cancer [15], and monitoring hosts in epidemiology [16]. The continuous evolution of machine learning is evident with the emergence of disruptive algorithms like transformers and their extensions, grounded in large language models and generative approaches [17,18]. While these advancements significantly enhance machine learning's impact across diverse fields, it's crucial to recognize that not all scenarios benefit from advanced algorithms. For instance, neural network algorithms, while capable of modeling performance in many cases, pose challenges due to their high data and computational resource requirements [19], coupled with the intricacies involved in model interpretation. On the contrary, traditional machine learning algorithms offer a pragmatic solution by being both convenient and fast, delivering commendable performance while remaining interpretable. Importantly, these algorithms don't necessitate extensive data and computational resources.

In light of these considerations, our study delved into the realm of traditional machine learning algorithms, including random forest (RF), logistic regression (LR), and support vector machine (SVM), to classify GBS host and explore potential cross-species transmission. Moreover, we investigated the interpretability of these models in more depth through the SHapley Additive exPlanations (SHAP) method, and identified key factors that are relevant to the hosts.

2. Methods

2.1. Isolates collection and whole genome sequencing

The study included genome sequence data from 1284 single host isolates. Among them, 486 were isolated locally from a tertiary hospital (n = 292) and local wet markets (n = 194) in Hong Kong as previously described [8], with contigs deposited into NCBI Bioprojects (PRJNA752017, PRJNA844521 and PRJNA844522), while the remaining data was from our previous study [20]. Briefly, DNA extraction was performed using Wizard Genome DNA Purification Kit (Promega, Madison, WI, USA), followed by library preparation using Nextera XT Library Preparation (Illumina, San Diego, CA, USA) or Riptide High Throughput Rapid Library Preparation Kit (Twist Bioscience, San Francisco, CA, USA) according to manufacturer's protocol. Genome sequencing was performed with NextSeq mid-output 500 system (Illumina, San Diego, CA, USA) to obtain an approximate minimum of 30 x average coverage of 150 bp paired-end sequence data. Genomes were assembled as previously described [8]. Quality control of reads was performed with FastQC prior to assembly with SPAdes assembler (v 3.5.0) [21]. Contigs of at least 500 bp were used for variant calling. SNP calling after assembly was conducted by Snippy (github.com/tseemann/snippy) workflow to reference genome *Streptococcus agalactiae* 2603V/R (Genbank ID: NC_004116) with default parameters [22]. This workflow contains SNP calling and annotation using freebayes and snpEff.

2.2. SNPs data processing for machine learning

We then transformed the SNPs data into the input data for machine learning following the procedure from our previous study [23,24]. We extracted the reference, variant alleles, and corresponding positional information from the raw variants data. We then merged the information from all isolates and labeled the loci with no variant information as N. The SNP matrix was then encoded by label encoding, where the A, C, G, T, and N in the SNP matrix were assigned to 1, 2, 3, 4, and 0. We discarded the locus that contained label 0 more than half of the sample size, which was considered not more reliable. In addition, we did not consider isolates belonging to fewer than 20 hosts or lacking data sources. Because such category labels are too few to be meaningful for machine learning. Thus, the final SNP matrix has 3895 columns and 1241 rows (corresponding to 1241 isolates and 3895 variants). Different host sources (Human, Bovine, Fish, and Pig) were used as class labels.

2.3. Construction and training of ML models

We randomly split the input data into training and test data, representing 80% (994 isolates) and 20% (247 isolates) of the data, respectively. We then constructed RF model, a LR model, and a SVM model using the caret R package v. 6.0–94 [25]. In the context of the random forest classifier, our training approach involved utilizing the rf method with mtry = 88 in the caret package. Conversely, for logistic regression, we opted for a penalized multinomial regression, employing the multinom method with a decay parameter set to 0.1 [26]. When it comes to the support vector machine, we employed polynomial kernels in implementing SVM, specifying the method as svmPoly and setting the degree to 1, scale to 10, and C to 1. During the training process, we facilitated multi-class classification by configuring the 'summaryFunction' parameter as 'multiClassSummary.' To comprehensively assess model performance, we employed a 10-fold cross-validation with 3 repeats. This was achieved by setting the 'method' parameter in the trainControl function to 'repeatedcv.' Additionally, to capture predictions for optimal tuning parameters, we set 'savePredictions = final'.

2.4. Evaluation of ML models

After model training, we assessed the performance of the three models on the independent test data via different evaluation metrics, including accuracy, precision, recall, sensitivity, specificity, F1 score, Kappa score [27], and Area Under the ROC (Receiver Operating Characteristics) curve (AUC) using 'multiClassSummary' [25,28,29]. For binary classification, their formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Here, TP represents True Positives (correctly predicted positive values), TN represents True Negatives (accurately predicted negative values), FN represents False Negatives (incorrectly predicted negative instead of

positive), and FP represents False Positives (incorrectly predicted positive instead of negative).

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

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

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specifity = \frac{TN}{TN + FP} \tag{6}$$

$$Kappa_{score} = 2 * \frac{TP * TN - FN * FP}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)} \tag{7}$$

While for multi-class classification, we calculated these metrics based on "one-versus-all" [25]. For example, the sensitivity of the human class is calculated against all the samples in the Bovine, Pig, and Fish classes. The function 'multiClassSummary' computes the overall accuracy and Kappa score using the predicted classes, and gives averages of the "one-versus-all" metrics such as precision, recall, sensitivity, specificity, and F1.

2.5. Interpretable machine learning by Shapley values

To improve the interpretability of our robust models and gain deeper insights into feature contributions, we employed SHAP (Shapley Additive exPlanations), which use game theory to assign credit for a model's prediction to each feature or feature value, to elucidate the model outputs [30–32]. Here, We calculated the Shapley values on the test data using SHAP package [30–32]. Specifically, we utilized the TreeExplainer algorithm from the SHAP package to interpret predictions from the RF model, while the LinearExplainer algorithm was employed for the LR model. We computed both the average Shapley value across all test data and the Shapley values for each class, including individual Shapley values for misclassified samples.

2.6. Correlation analysis of misclassified samples

To further explain the incorrect predictions made by our high-performing models, we isolated the misclassified sample data from the SNP matrix. Subsequently, we conducted a Pearson correlation analysis using the corrplot R package (version 0.92) to discern potential patterns or relationships.

2.7. Phylogenetic analysis of misclassified samples

Phylogenetic trees serve as powerful tools to illustrate relationships between species. To facilitate the understanding of misclassified samples, we performed a phylogenetic tree analysis based on SNP information using VCF2PopTree [33]. We used it to calculate the genetic distance for each SNP and then summed up the obtained distances to get the total number of differences for the whole genome. The pairwise matrix of these differences can be used to construct a phylogeny. Subsequently, we exported the Newick tree format and visualized it using ggtree [34].

2.8. Variants annotation

We utilized the Ensembl Variant Effect Predictor (VEP) web interface [35] to annotate variants and identify the associated genes corresponding to previously identified SNPs. The reference genome we used for annotation is *Streptococcus agalactiae* 2603V/R (GCA_000007265). The web portal accessed for annotation is https://bacteria.ensembl.org/Streptococcus_agalactiae_2603v_r_gca_000007265/Tools/VEP, with default parameters applied. We then used maftools to visualize the annotated results [36].

2.9. Probability calibration for multi-class classification

Probability calibration is a method used to align the predicted probabilities generated by a model with their actual probabilities. Here we used calibration_curve function from scikit learn module [37].

2.10. Statistical methods

To assess and compare the performance of various models, we conducted pairwise comparisons (RF-vs-LR, RF-vs-SVM, SVM-vs-LR) using the Wilcoxon test with 'method = 'wilcox.test', a non-parametric method, in R ggpubr package [38,39]. To correct for multiple comparisons, we applied the Bonferroni-Sidak adjustment method by setting p.adjust.method = 'bonferroni'. We also calculated the standard errors (se) by calculating the averages of metrics over the 10-folds cross-validation in 3 repeats and added error bar (±se) on the top of bar plot using R ggpubr package.

3. Results

3.1. Origin and molecular characteristics of GBS data

We collected a total of 1284 GBS isolates, with 836 isolates sourced from human, 186 from bovine, 172 from fish, and 47 isolates were from pigs (Table 1). Besides, 37 isolates were collected from various other hosts, each with fewer than 20 representatives (labeled as 'Others' in Table 1), while six isolates lacked identifiable data sources (denoted as 'Null' in Table 1). We then analyzed the characteristics of serotypes, sequence type (ST), and clonal complex (CC) of GBS across different hosts, which were classically used to describe the molecular epidemiology of bacterial populations and in the multilocus sequence typing system (MLST) for GBS [40].

The predominant capsular serotypes observed in our dataset included III (n = 324), Ia (n = 313), II (n = 234), V (n = 153), IV (n = 98), and Ib (n = 91) (Table 1). Among human-derived GBS strains, the prevalent serotypes comprised III, V, Ia, IV, II, and Ib, accounting for 94.4% of the total (Fig. 1a), similar to previously reported proportions [30–32]. Notably, the primary serotypes identified in GBS from bovines, fish, and pigs were type II, Ia, and III, respectively, with proportions of 83.3%, 95.9%, and 93.6% (Fig. 1a). These findings suggest shared serotype characteristics among GBS from humans, pigs, and fish, challenging straightforward host classification based on serotype alone.

Regarding STs, GBS exhibited a diverse distribution, encompassing ST7 (n = 157), ST17 (n = 130), ST1 (n = 124), ST23 (n = 99), ST61 (n = 88), ST19 (n = 84), ST12 (63), ST651 (n = 44), and other types (n = 539) (Table 1). Specifically, human-derived GBS predominantly comprised ST17, ST1, ST23, ST19, and ST12, totaling 56.8%. Conversely, GBS from bovines, fish, and pigs were primarily represented by ST61, ST7, and ST651, constituting proportions of 46.8%, 87.8%, and 85.1%, respectively (Fig. 1b). Despite certain GBS lineages displaying strong host associations, our results [41], revealed diversity in ST distributions across hosts, particularly evident in humans and pigs, indicating limitations in host classification solely based on STs.

Similarly, distinct host origins harbored overlapping CCs, with human-derived GBS mainly featuring CC1, CC17, CC19, CC23, and CC12, while bovine GBS were predominantly represented by CC17, CC67, and CC1. Meanwhile, fish-derived GBS were predominantly CC1, and pig-derived GBS were mainly CC103 (Fig. 1c). Thus, CCs also proved inadequate for precise host classification of GBS. Furthermore, we observed associations between serotypes and CCs; for instance, Serotype

Table 1
Overview of the number of each characteristic in the dataset.

Characteristics	All Isolates (n = 1284)								
Isolation Host	**Human**	**Bovine**	**Fish**	**Pig**	**Others**	**Null**			
	836	186	172	47	37	6			
Capsular Serotype	**Ia**	**Ib**	**II**	**III**	**IV**	**V**	**Others**	**Null**	
	313	91	234	324	98	153	70	1	
Sequence Type (ST)	**ST7**	**ST17**	**ST1**	**ST23**	**ST61**	**ST19**	**ST12**	**ST651**	**Others**
	157	130	124	99	88	84	63	44	539
Clonal Complex (CC)	**CC1**	**CC17**	**CC19**	**CC23**	**CC12**	**CC67**	**CC103**	**Others**	
	353	241	129	122	80	70	63	226	

Isolates with fewer than 20 characteristic types within each host were named as 'Others', and isolates lacking specific characteristics data were named as 'Null'. These data are not considered in the following analyses.

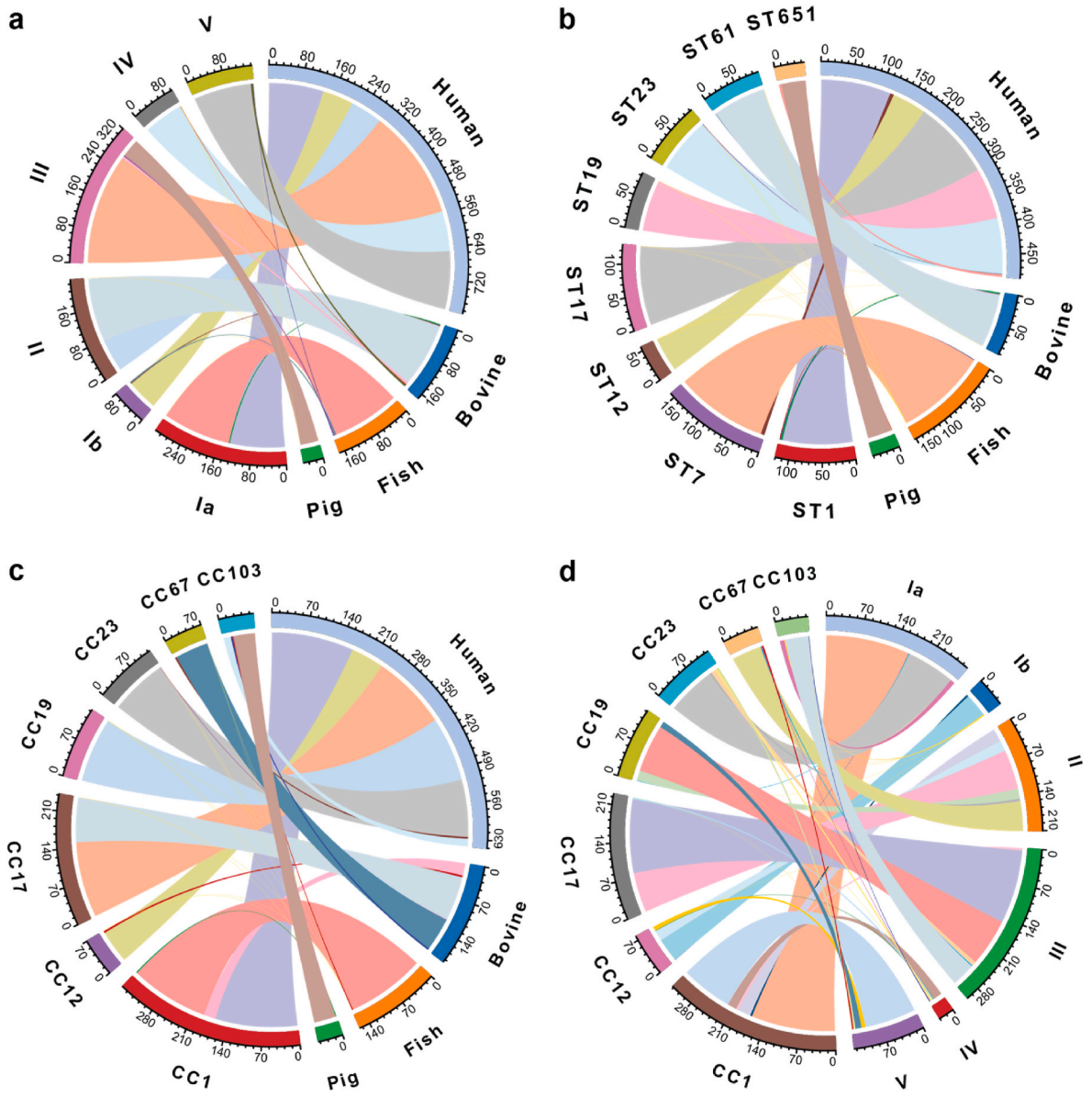

Fig. 1. Data characteristics. **a-c** Number of **(a)** capsular serotypes, **(b)** sequence type (ST), and **(c)** clonal complex (CC) associated with each host of GBS isolates. **d** Correlation of the number of capsular serotypes, sequence type, and clonal complex of GBS isolates. Isolates that fell in to 'Others' group in Table 1 are not shown in the figures.

Ia correlated primarily with CC1 and CC23, Ib with CC12, III with CC17, CC19, and CC103, and V with CC1 (Fig. 1d).

In summary, our analysis provides insights into the host origins of GBS, elucidating their molecular characteristics and associations. However, neither serotype, ST, nor CC singularly represents the optimal choice for host classification.

3.2. ML models based on genome-wide mutation yield superior predictions for GBS hosts

To accurately predict GBS host without relying on specific serotype, ST, or CC information, we developed three ML models, namely RF, LR, and SVM, leveraging genome-wide mutations. Unlike conventional approaches that filter variants based on prior biological knowledge, we utilized the entire spectrum of genome-wide variants, resulting in a final matrix comprising 3895 variants (see Methods for detailed procedures). Upon training the models, we observed impressive accuracy scores for RF and LR on the training dataset, achieving accuracies of 0.97 ± 0.003 and 0.95 ± 0.004, respectively (Fig. 2a, Table S1). Additionally, evaluation metrics such as the AUC and F1 score exceeded 0.90, with the RF model reaching 0.99 ± 0.002 for AUC and 0.95 ± 0.005 for F1, and the LR model achieving 0.98 ± 0.004 for AUC and 0.93 ± 0.007 for F1. Conversely, the performance of the SVM model varied across different evaluation metrics, generally lagging behind RF and LR models (Fig. 2a).

On the test set, both RF and LR models exhibited outstanding performance, surpassing SVM in predicting GBS host (Fig. 2b). Specifically, the F1 score for both RF and LR models, reached 0.98, while the SVM was 0.86 for human prediction (Table S2).

On the other hand, it's essential to consider the potential for model overfitting. To address this concern, we conducted model calibration plots to see the difference between predicted and true probabilities. Due to the complexity of multiclassification models, our results vary widely in each class label (Fig. S1). In general, the results remain acceptable. Overall, RF and LR models consistently demonstrated accurate prediction of GBS hosts on both the training and test datasets.

3.3. Analysis of feature contributions by Shapley values

Feature importance scores are commonly calculated as one of the metrics to explain the impact of features on model performance, which provides the global contribution of features to model output. However, they often lack granularity, providing no insight into the impact on individual observations or impact of positive/negative direction [42–45]. Shapley Additive exPlanations (SHAP) presents a solution to this limitation [30]. SHAP values can explain the contribution of each feature to the model output at both the global dataset and at the per-sample level, and can also provide the positive or negative impact of each feature on the model [30]. Thus, to delve deeper into the interpretations of our top-performing models (RF and LR), we computed the SHAP values on the test dataset. We identified the 20 most influential features with the highest average SHAP values for both RF and LR models across all host classes (Fig. 3a and b). Notably, the impact of features can vary across different host classes. Hence, we further computed the SHAP value on each data point for each host class of both models (Fig. 3c and d). In the distribution plots, the Y-axis in the distribution plots shows the 20

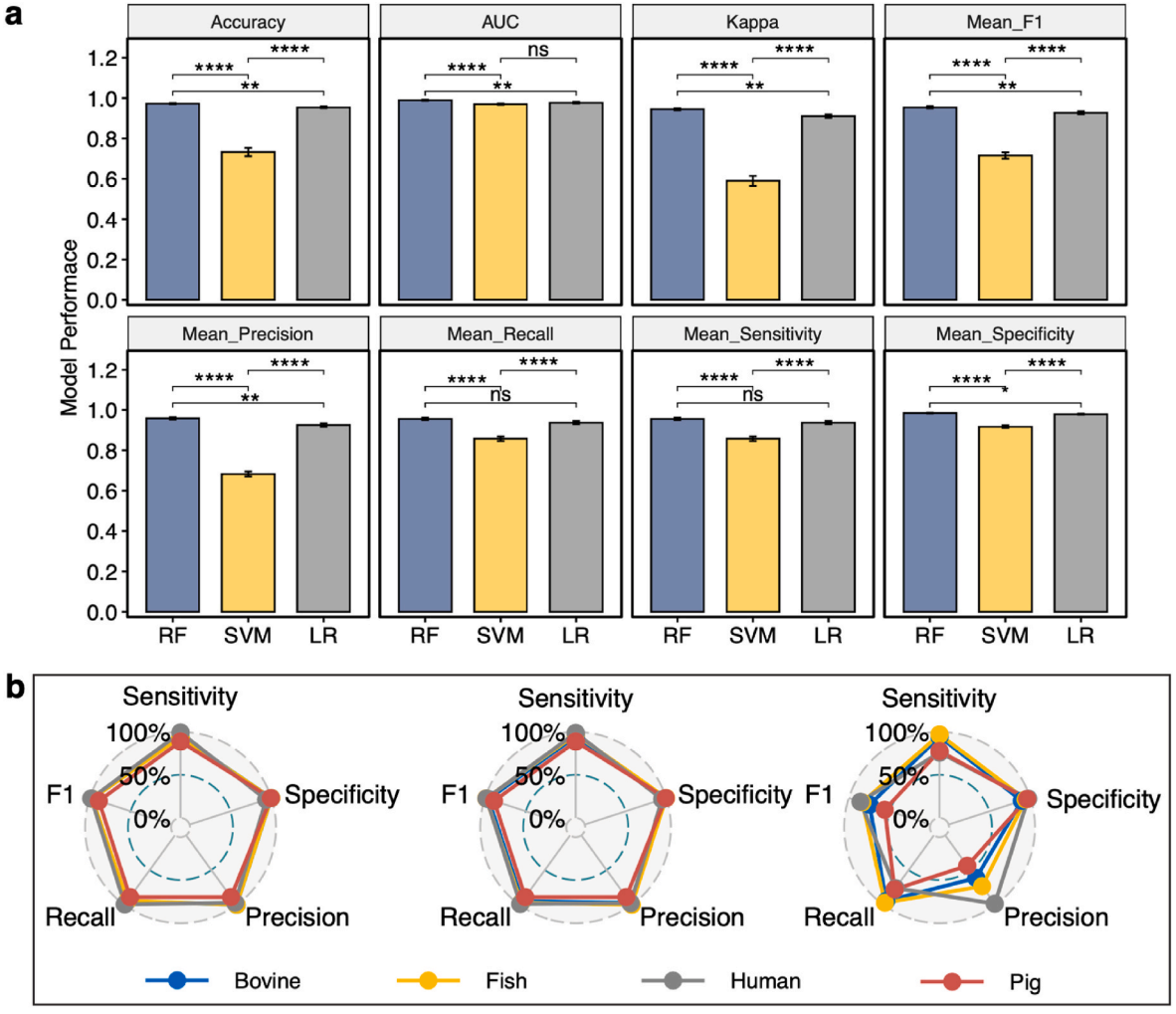

Fig. 2. Performance of ML models (RF, LR, SVM) for predicting GBS hosts. **a** Models performance on the training dataset as defined by accuracy, AUC of ROC, Kappa, F1 score, Precision, Recall, Sensitivity, and Specificity. **b** The predictive performance of ML models (RF, LR, and SVM from left to right) for each host was evaluated on the test dataset using metrics such as F1 score, Precision, Recall, Sensitivity, and Specificity. Statistical comparisons were performed using the Wilcoxon test. *$p <$ 0.05; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.

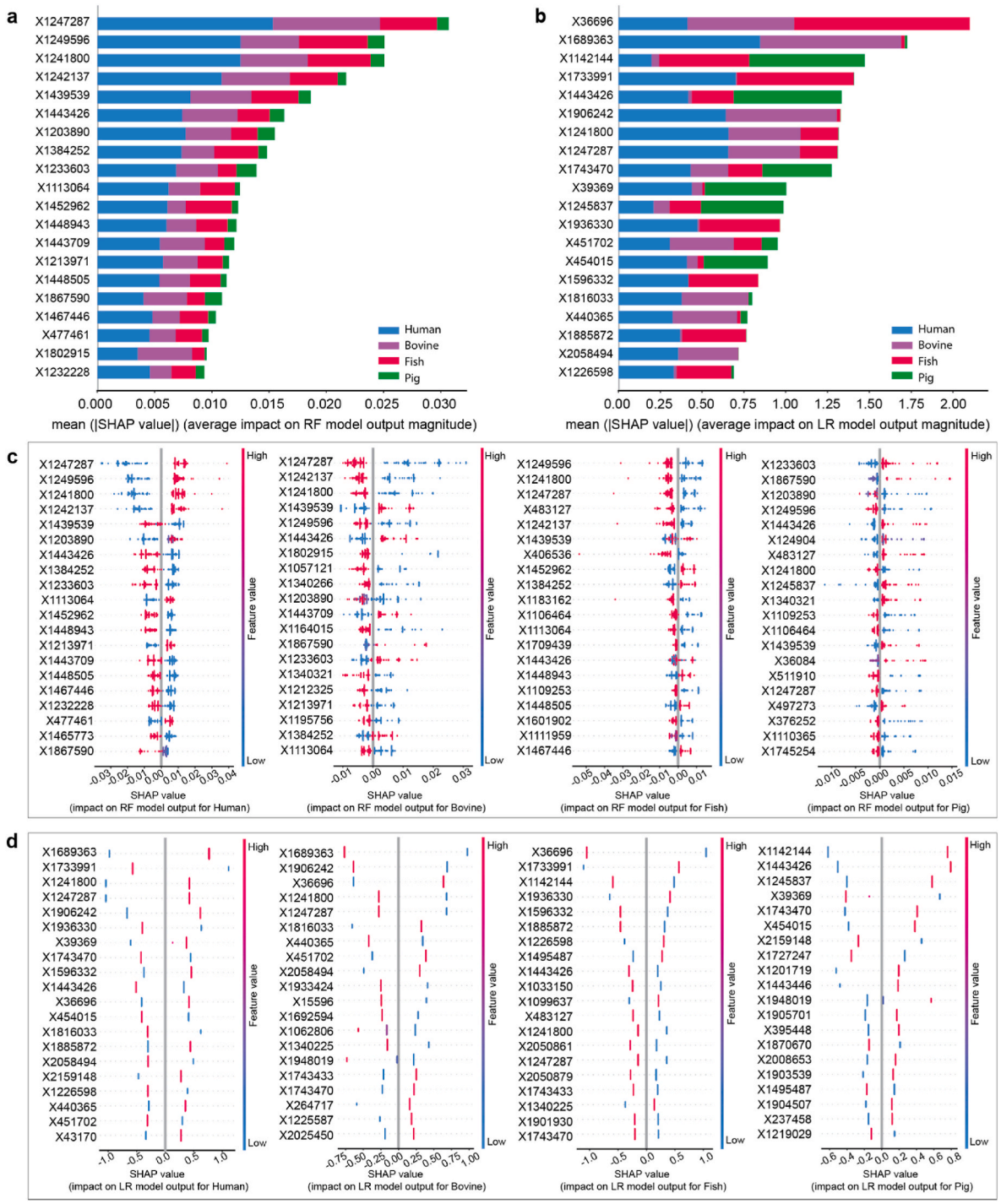

Fig. 3. Quantification of feature impact on RF and LR model predictions by analyzing SHAP values. **a-b** The average impact of the SNPs on hosts classification for RF (**a**) and LR (**b**) models based on mean SHAP value. The Y-axis indicates the 20 most influential SNPs, and the number is the position of SNP. The X-axis indicates the average impact of each SNP on the RF model output magnitude, and the color bar indicates the category of the host. **c-d** Impact of the 20 most influential SNPs of RF (**c**) and LR (**d**) models on each of the four hosts. The scatter plot shows the distribution of SHAP values in all test samples. Y-axis indicates top 20 influential SNPs, and the number is the position of SNPs. The colors of the scatter plot indicate the feature value, from blue to red, representing low feature values to high feature values according to the color bar on the right.

features exerting the highest influence on predictions for each host class. Meanwhile, the X-axis illustrates the positive and negative influence of feature values' magnitude (from blue to red, representing low feature values to high feature values according to the color bar on the right). In summary, this analysis discerned the most influential features for the overall GBS host classification and identified distinct influential features for each host class within RF and LR models.

Further insight into GBS transmission among hosts through misclassification analysis.

On the test data, our RF model misclassified six samples, while the LR model misclassified five samples, with an overlap of five misclassified samples between the two models. The true and predicted labels for these samples are detailed in Table 2. Notably, five samples (Fish1, Pig, Bovine1, Bovine2, Fish2) were misclassified as Human, while one

Table 2
True and predicted class of misclassified samples.

Sample ID	Truth	Prediction
Fish1	Fish	Human
Pig	Pig	Human
Bovine1	Bovine	Human
Bovine2	Bovine	Human
Fish2	Fish	Human
Human	Human	Pig

(Human) was misclassified as Pig.

To explore the potential for horizontal interspecies transmission underlying these misclassifications, we conducted a correlation analysis on these six misclassified samples. We observed a strong correlation between human and pig samples, with a coefficient of 0.91 (Fig. 4a). Additionally, a notable correlation of 0.71 was found between Bovine2 and Fish2 samples (Fig. 4a). We also analyzed the phylogenetic relationships between them based on mutation variants. The results showed that samples Human and Pig have a close evolutionary relationship, as well as the samples Bovine2 and Fish2 (Fig. 4b). These results suggest high similarity of variants information in the samples Bovine, Fish, and Pig. It is possible that certain variants result in inter-host-species transmission between fish and bovine, as well as pigs and human.

Samples (Fish1, Pig, Bovine1, Bovine2, Human) were misclassified by both RF and LR models. The sample (Fish2) was misclassified by the RF model.

Hence, to further determine which features had a greater impact on the misclassified samples, we calculated the Shapley values in each misclassified sample for both RF (Fig. 5) and LR models (Fig. 6).

3.4. Annotation of potentially influential variants

Based on our findings, we identified a total of 109 SNPs that could

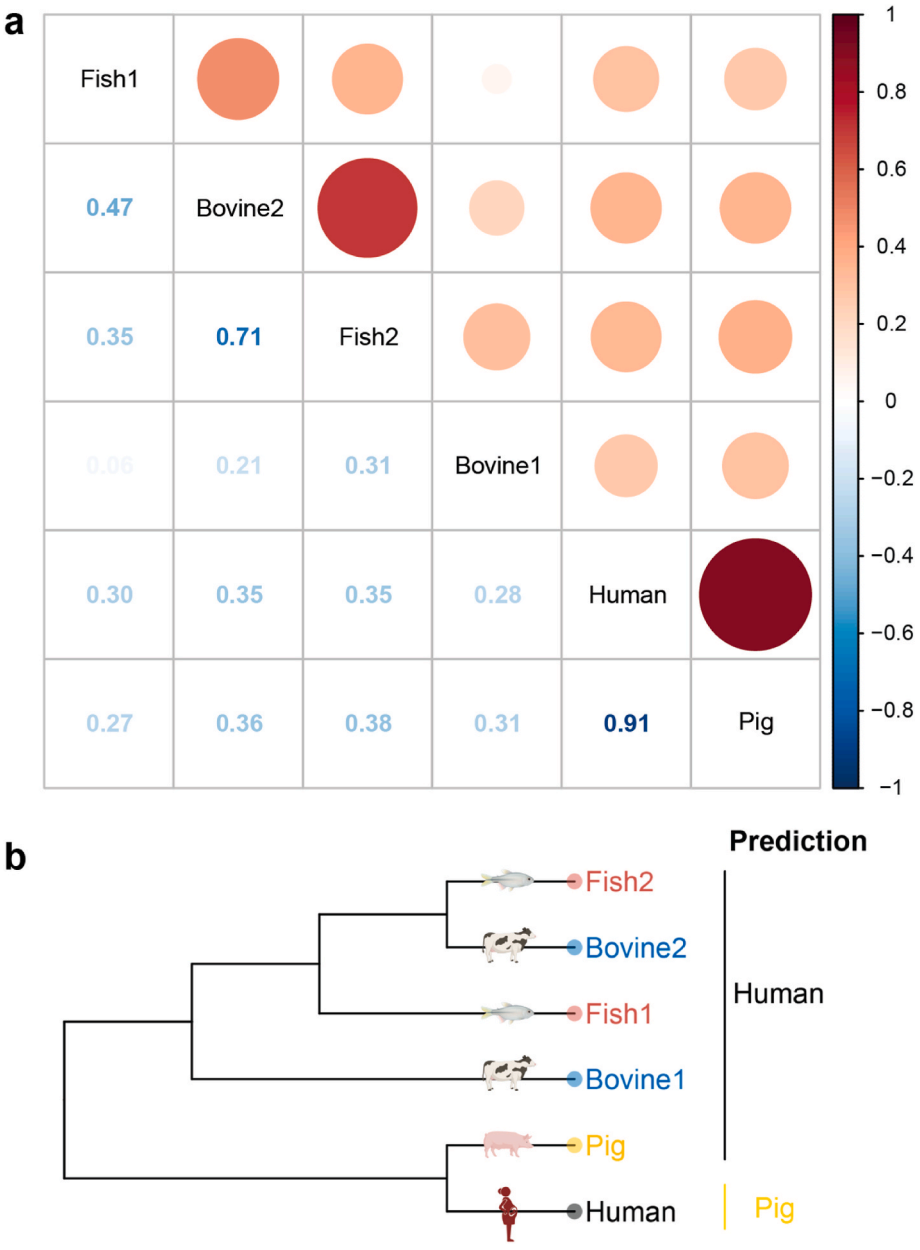

Fig. 4. Analysis of misclassification samples. **a** Correlations among all misclassified samples. **b** Phylogenetic tree of all misclassified samples. Node labels are based on true labels. The labels in the right panel are the results of the false predictions from RF and LR models. Some icons were created with BioRender.com.

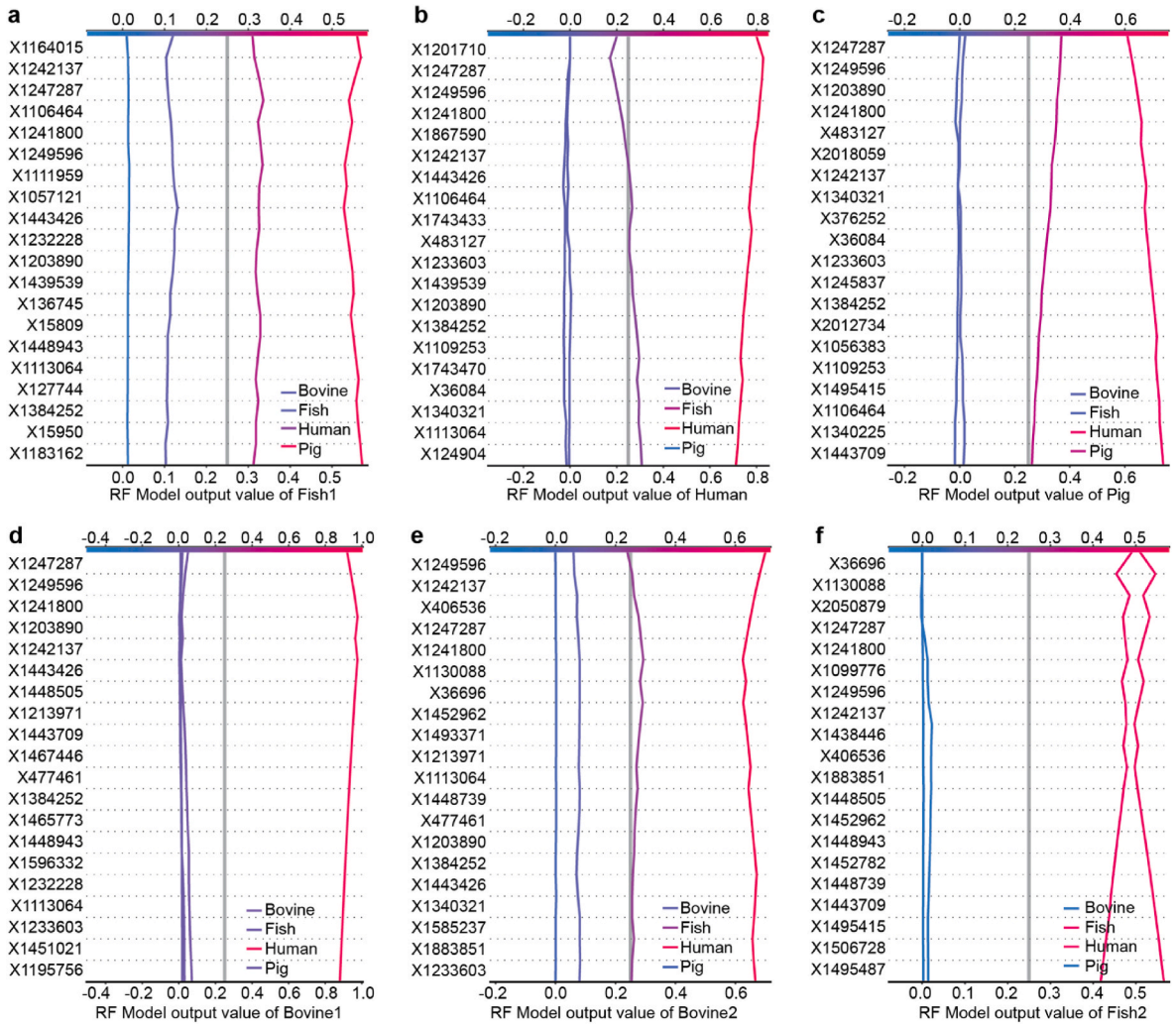

Fig. 5. SNPs impact of RF model on hosts classification in each misclassified sample. **a-f** Impact of the 20 most influential SNPs of RF model in misclassified samples including Fish1 **(a)**, Human **(b)**, Pig **(c)**, Bovine1 **(d)**, Bovine2 **(e),** and Fish2 **(f)**.

potentially influence GBS host classification or interspecies transmission. We then analyzed the functional implications of these SNPs using VEP annotation tools (for details see Methods). This analysis yielded 1110 annotations for the 109 SNPs (Table S3), including various types such as upstream and downstream gene variants, synonymous variants, missense variants, and coding sequence variants, which account for 47.58 %, 44.95%, 4.77%, 2.34%, 0.36%, respectively (Fig. 7a).

To focus our investigation, we retained only the SNPs located within genes, excluding those annotated upstream or downstream. The filtered list comprised synonymous variants, missense variants, and coding sequence variants, accounting for 60.00%, 33.00%, and 7.00%, respectively (as depicted in Fig. 7a). Additionally, we examined the transition and transversion percentages among the filtered variants, with C > T and T > C transitions representing a larger proportion (as indicated in Fig. 7b).

Furthermore, we specifically analyzed the missense variants and visualized their distribution across genes using an oncoplot (as shown in Fig. 7c), providing insight into their relevance within each classification category.

4. Discussion and conclusions

Host prediction provides cues to the initial contact point for local outbreaks in an effort to contain the spread within the broader community. However, unlike some bacteria like *Staphylococcus aureus*, where specific associations between hosts and STs are well-documented, such information is relatively limited for GBS. ML methods for host prediction have been more established and documented for parasites and viruses [46–51]. In this study, we employed three ML models (RF, LR, and SVM) based on genome-wide mutations to predict GBS hosts, achieving high accuracy, especially with RF and LR models, reaching the accuracies of 0.97 ± 0.003 and 0.95 ± 0.004, respectively. Our methods overcome the limitations of classification based on molecular characteristics, such as ST, CC, and capsular serotypes [52]. Moreover, we provided the most influential SNPs for predicting GBS hosts and related genes of these SNPs. In future studies, we can experimentally validate these most influential SNPs and genes and then design rapid detection kits for rapid detection and timely control of the spread of GBS. This prediction could also deduce host dependency factors and host-pathogen protein interactions, which can be used as targets for rapid detection tests [53,54].

Additionally, the zoonotic potential of the bacteria has been demonstrated in previous outbreaks in the community and fish farms [11,55]. Our analysis of misclassified samples also revealed possible cross-species transmission between humans and fish, bovine, and pigs. On the other hand, these findings may also indicate human-animal contact which may have contracted the bacteria and lead to zoonosis. This is in line with a previous GWAS study indicating certain GBS lineage exhibit host-specificity, while some may be host generalists, in which the latter may have a possibility of recombination with other host

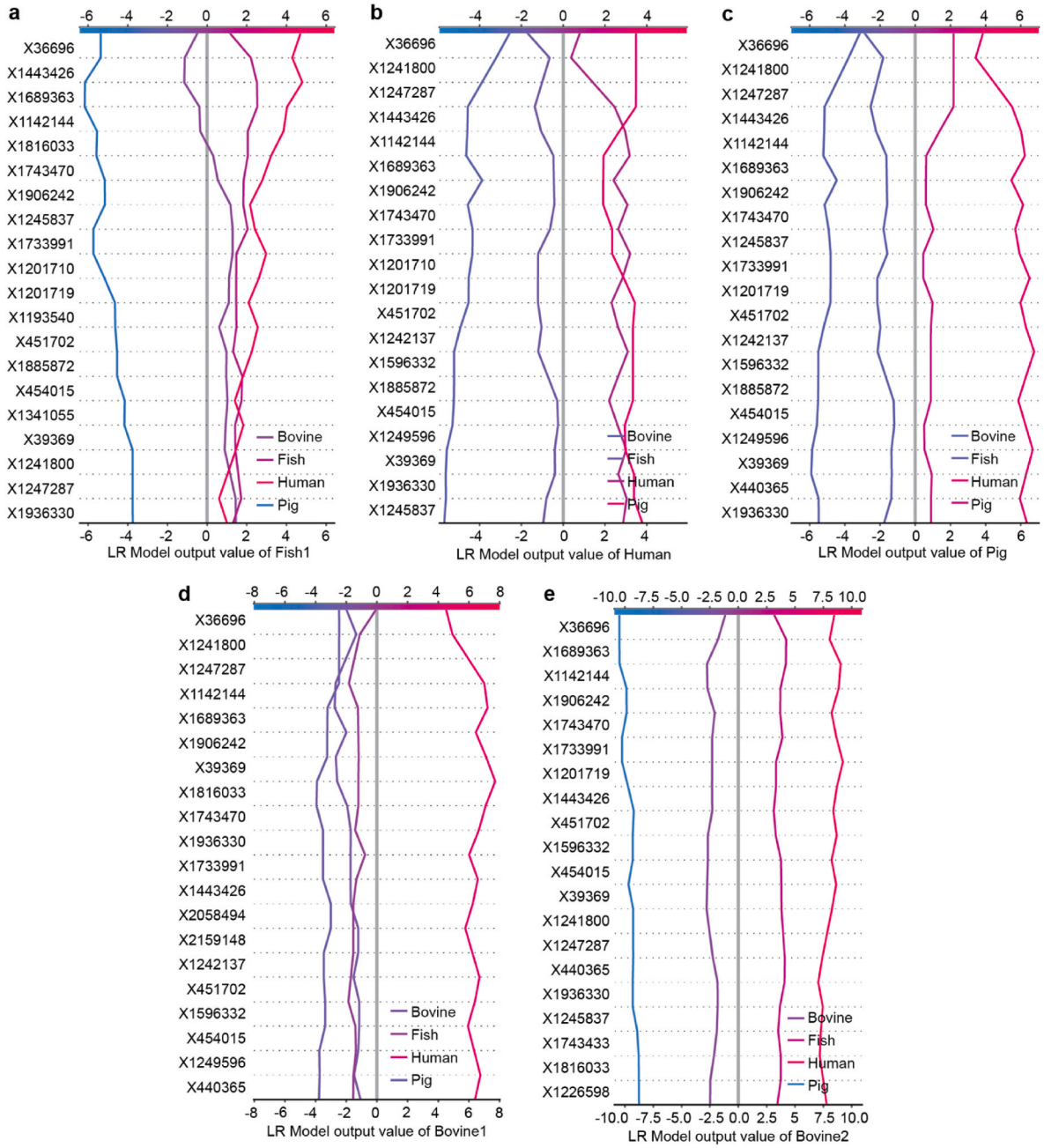

Fig. 6. SNPs impact of LR model on hosts classification in each misclassified sample. **a-e** Impact of the 20 most influential SNPs of RF model in misclassified samples including Fish1 **(a)**, Human **(b)**, Pig **(c)**, Bovine1 **(d)**, and Bovine2 **(e)**.

strains [41]. Moreover, Crestani et al. [41] reported that adaptation of GBS in humans, bovine, and fish was associated with C5a-peptidase (scpB, SAG1594), we also identified one mutant locus 1596332 annotated as an upstream variant of scpB (SAG1594, see in Table S3). Our results indicated that this mutation significantly influences the classification of human and fish (Fig. 3d). Additionally, it is a critical locus for the misclassification between pig and human (Fig. 6b and c), as well as between bovine and human (Fig. 6d and e).

With our model, we are able to deduce host source of GBS strains, which could, in turn, thereby facilitating investigative efforts in public health and enabling the implementation of control measures from a One Health perspective. Furthermore, this approach can elucidate zoonotic characteristics of GBS and shed light on potential instances of reverse zoonosis within the evolutionary trajectory of the GBS population.

A extended application of our models on GBS host prediction also includes the selection of appropriate antibiotics therapy, thus circumventing antibiotic resistance if the host source is known. Importantly, rigorous laboratory testing is necessary to further explore the potential and limitations of this approach for host prediction, especially for generalized application in public health.

In summary, we have developed three ML models (RF, LR, and SVM) to predict the broad host origin of GBS. Among them, RF and LR models were demonstrated to be robust and effective. Furthermore, the ML interpretability analysis of these models provided valuable insights into the contribution of different features, offering a diverse range of targets with potential applications. Our study greatly enhances the ability to track potential GBS hosts, enabling more targeted field sampling for specific host species and optimized surveillance. This approach helps to recognize and distinguish between our reservoir of GBS in the human and animal host, and implement measures that may prevent the spread

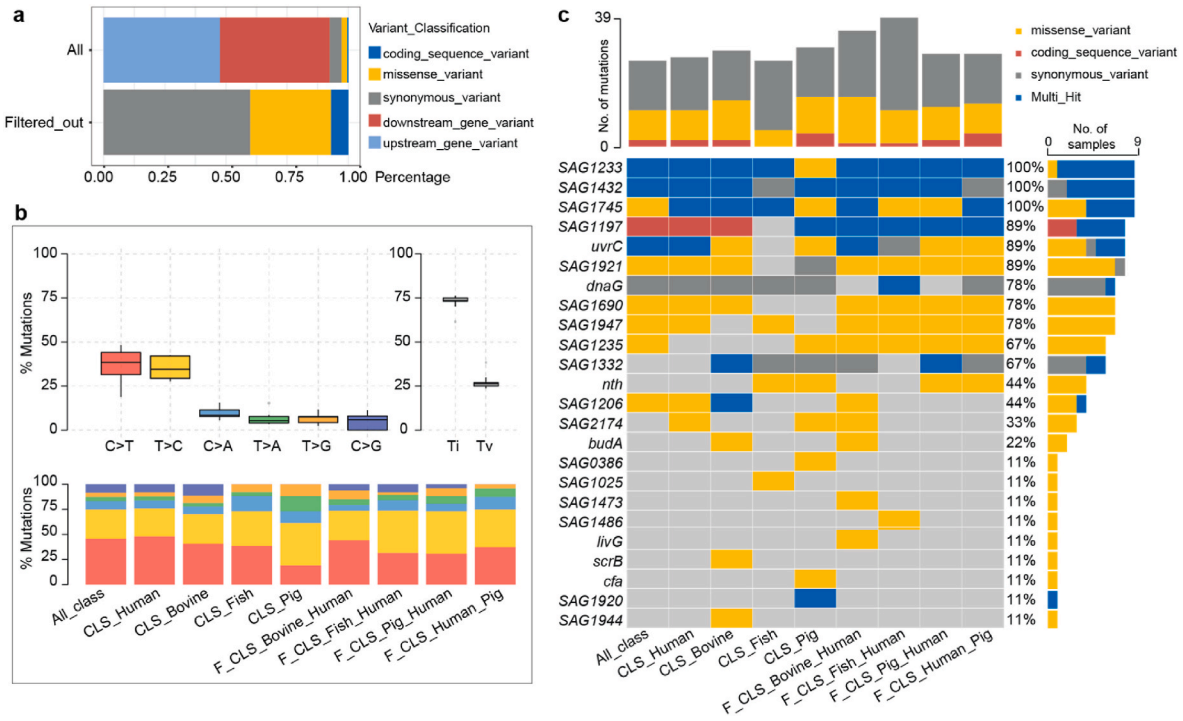

Fig. 7. Annotation of influential variants. **a** Variant classification types of all influential variants and filtered variants based on previous results. **b** The percentage of transition and transversion on filtered variants. **c** The oncoplot shows the missense variants and related genes.

of GBS between humans and animals, thereby reducing public health risks.

Availability of data and materials

The code and data used in this work are available at Github, and the additional files can also be found at Github https://github.com/YunxiaoRen/Hosts_Classification_of_GBS.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

CRediT authorship contribution statement

Yunxiao Ren: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Carmen Li:** Resources. **Dulmini Nanayakkara Sapugahawatte:** Resources. **Chendi Zhu:** Resources. **Sebastian Spänig:** Software. **Dorota Jamrozy:** Resources. **Julian Rothen:** Resources. **Claudia A. Daubenberger:** Resources. **Stephen D. Bentley:** Resources. **Margaret Ip:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Dominik Heider:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements/Funding

We acknowledge the funding support provided by the Federal Ministry of Education and Research (BMBF) grant ID 57513593, Deep-iAMR (FKZ 031L0209B), German Academic Exchange Service (DAAD) and Research Grants Council (RGC) Joint Research Scheme [PIs to DH (Germany) and MI (Hong Kong) grant no: G-CUHK403/21]. The project was partially supported by the Food and Health Bureau, Government of Hong Kong Special Administrative Region under the Health and Medical Research Fund (HMRF, Grant No.#17160212, PI: MI). We also acknowledge the partial support from the Juno project funded by the Bill and Melinda Gates Foundation (grant code INV-010426).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108185.

Abbreviations

GBS	Group B *Streptococcus*
RF	Random forest
LR	Logistic regression
SVM	Support vector machine
ShAP	SHapley Additive exPlanations
SNP	Single nucleotide polymorphism
ROC	Receiver operating characteristics
AUC	Area under the ROC curve
VEP	Variant effect predictor
ST	Sequence type
CC	Clonal complex

References

[1] K. Le Doare, P.T. Heath, An overview of global GBS epidemiology, Vaccine 31 (Suppl 4) (2013) D7–D12, https://doi.org/10.1016/j.vaccine.2013.01.009.

[2] Prevention of Perinatal Group B Streptococcal Disease, (n.d.). https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5910a1.htm (accessed June 9, 2023).

[3] G. Kwatra, M.C. Cunnington, E. Merrall, P.V. Adrian, M. Ip, K.P. Klugman, W. H. Tam, S.A. Madhi, Prevalence of maternal colonisation with group B streptococcus: a systematic review and meta-analysis, Lancet Infect. Dis. 16 (2016) 1076–1084, https://doi.org/10.1016/S1473-3099(16)30055-X.

[4] A. Almeida, I. Rosinski-Chupin, C. Plainvert, P.-E. Douarre, M.J. Borrego, C. Poyart, P. Glaser, Parallel evolution of group B Streptococcus hypervirulent clonal complex 17 unveils new pathoadaptive mutations, mSystems 2 (2017) e00074, https://doi.org/10.1128/mSystems.00074-17, 17.

[5] Prevention of Group B Streptococcal Early-Onset Disease in Newborns, (n.d.). https://www.acog.org/en/clinical/clinical-guidance/committee-opinion/articles/2020/02/prevention-of-group-b-streptococcal-early-onset-disease-in-newborns (accessed June 9, 2023).

[6] A.C. Seale, F. Bianchi-Jassir, N.J. Russell, M. Kohli-Lynch, C.J. Tann, J. Hall, L. Madrid, H. Blencowe, S. Cousens, C.J. Baker, L. Bartlett, C. Cutland, M. G. Gravett, P.T. Heath, M. Ip, K. Le Doare, S.A. Madhi, C.E. Rubens, S.K. Saha, S. J. Schrag, A. Sobanjo-Ter Meulen, J. Vekemans, J.E. Lawn, Estimates of the burden of group B streptococcal disease worldwide for pregnant women, stillbirths, and children, Clin. Infect. Dis. 65 (2017) S200–S219, https://doi.org/10.1093/cid/cix664.

[7] A. Navarro-Torné, D. Curcio, J.C. Moïsi, L. Jodar, Burden of invasive group B Streptococcus disease in non-pregnant adults: a systematic review and meta-analysis, PLoS One 16 (2021) e0258030, https://doi.org/10.1371/journal.pone.0258030.

[8] D.N. Sapugahawatte, C. Li, P. Dharmaratne, C. Zhu, Y.K. Yeoh, J. Yang, N.W.S. Lo, K.T. Wong, M. Ip, Prevalence and characteristics of Streptococcus agalactiae from freshwater fish and pork in Hong Kong wet markets, Antibiotics 11 (2022) 397, https://doi.org/10.3390/antibiotics11030397.

[9] K.A. Anderson, A.M. Schaefer, C.D. Rice, Quantifying circulating antibody activities against the emerging environmental pathogen, Streptococcus agalactiae, in wild captured bull sharks, spotted eagle rays, bottlenose dolphins, and loggerhead turtles, Fish Shellfish Immunol. Rep. 2 (2021) 100024, https://doi.org/10.1016/j.fsirep.2021.100024.

[10] L.C. Simões, F.G. Fernandes, I.C.M. de Oliveira, A.B. de Almeida Corrêa, N.S. Costa, L.M.A. Oliveira, A.C.N. Botelho, S.E.L. Fracalanzza, L.M. Teixeira, T.C.A. Pinto, Characteristics of Streptococcus agalactiae belonging to CC103 clone circulating among dairy herds and pregnant women in Brazil, Braz. J. Microbiol. (2023), https://doi.org/10.1007/s42770-023-01017-y.

[11] P. Rajendram, W. Mar Kyaw, Y.S. Leo, H. Ho, W.K. Chen, R. Lin, D.P. Pratim, H. Badaruddin, B. Ang, T. Barkham, A. Chow, Group B Streptococcus sequence type 283 disease linked to consumption of raw fish, Singapore, Emerg. Infect. Dis. 22 (2016) 1974–1977, https://doi.org/10.3201/eid2211.160252.

[12] A. Kaur, A.P.S. Chauhan, A.K. Aggarwal, Prediction of enhancers in DNA sequence data using a hybrid CNN-DLSTM model, IEEE ACM Trans. Comput. Biol. Bioinf 20 (2023) 1327–1336, https://doi.org/10.1109/TCBB.2022.3167090.

[13] A. Kaur, A.P.S. Chauhan, A.K. Aggarwal, Machine learning based comparative analysis of methods for enhancer prediction in genomic data, in: 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), 2019, pp. 142–145, https://doi.org/10.1109/ICCT46177.2019.8969054.

[14] A. Kaur, A.P.S. Chauhan, A.K. Aggarwal, Dynamic deep genomics sequence encoder for managed file transfer, IETE J. Res. 0 (2022) 1–13, https://doi.org/10.1080/03772063.2022.2060869.

[15] A. Kaur, A.P.S. Chauhan, A.K. Aggarwal, An automated slice sorting technique for multi-slice computed tomography liver cancer images using convolutional network, Expert Syst. Appl. 186 (2021) 115686, https://doi.org/10.1016/j.eswa.2021.115686.

[16] C.C.S. Tan, S.D. Lam, D. Richard, C.J. Owen, D. Berchtold, C. Orengo, M.S. Nair, S. V. Kuchipudi, V. Kapur, L. van Dorp, F. Balloux, Transmission of SARS-CoV-2 from humans to animals and potential host adaptation, Nat. Commun. 13 (2022) 2988, https://doi.org/10.1038/s41467-022-30698-6.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2017. http://arxiv.org/abs/1706.03762. (Accessed 6 December 2022).

[18] M.E. Consens, C. Dufault, M. Wainberg, D. Forster, M. Karimzadeh, H. Goodarzi, F. J. Theis, A. Moses, B. Wang, To Transformers and beyond: Large Language Models for the Genome, 2023, https://doi.org/10.48550/arXiv.2311.07621.

[19] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021) 53, https://doi.org/10.1186/s40537-021-00444-8.

[20] J. Rothen, J.F. Pothier, F. Foucault, J. Blom, D. Nanayakkara, C. Li, M. Ip, M. Tanner, G. Vogel, V. Pflüger, C.A. Daubenberger, Subspecies typing of Streptococcus agalactiae based on ribosomal subunit protein mass variation by MALDI-TOF MS, Front. Microbiol. 10 (2019), https://doi.org/10.3389/fmicb.2019.00471.

[21] A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, A. Korobeynikov, Using SPAdes de novo assembler, Curr. Protoc. Bioinf. 70 (2020) e102, https://doi.org/10.1002/cpbi.102.

[22] T. Seemann, Snippy, 2023. https://github.com/tseemann/snippy. (Accessed 22 May 2023).

[23] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, A.-C. Hauschild, O. Schwengers, D. Heider, Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning, Bioinformatics (2021) btab681, 10/gnpq2p.

[24] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, O. Schwengers, D. Heider, Deep transfer learning enables robust prediction of antimicrobial resistance for novel antibiotics, Antibiotics 11 (2022) 1611, https://doi.org/10.3390/antibiotics11111611.

[25] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Software 28 (2008) 1–26, https://doi.org/10.18637/jss.v028.i05.

[26] caret/models/files/glmnet.R at master · topepo/caret, GitHub (n.d.). https://github.com/topepo/caret/blob/master/models/files/glmnet.R (accessed January 25, 2024).

[27] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. 22 (2012) 276–282.

[28] M. H, M.N. S, A review on evaluation metrics for data classification evaluations, IJDKP 5 (2015) 1–11, https://doi.org/10.5121/ijdkp.2015.5201.

[29] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. 12 (2022) 5979, https://doi.org/10.1038/s41598-022-09954-8.

[30] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, in: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. (Accessed 27 April 2023).

[31] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.

[32] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D. E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (2018) 749–760, https://doi.org/10.1038/s41551-018-0304-0.

[33] S. Subramanian, U. Ramasamy, D. Chen, VCF2PopTree: a client-side software to construct population phylogeny from genome-wide SNPs, PeerJ 7 (2019) e8213, https://doi.org/10.7717/peerj.8213.

[34] G. Yu, D.K. Smith, H. Zhu, Y. Guan, T.T.-Y. Lam, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data, Methods Ecol. Evol. 8 (2017) 28–36, https://doi.org/10.1111/2041-210X.12628.

[35] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl variant effect predictor, Genome Biol. 17 (2016) 122, https://doi.org/10.1186/s13059-016-0974-4.

[36] A. Mayakonda, D.-C. Lin, Y. Assenov, C. Plass, H.P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer, Genome Res. 28 (2018) 1747–1756, https://doi.org/10.1101/gr.239244.118.

[37] Probability Calibration for 3-class Classification in Scikit Learn, GeeksforGeeks, 2023. https://www.geeksforgeeks.org/probability-calibration-for-3-class-classification-in-scikit-learn/. (Accessed 25 January 2024).

[38] ggplot2 Based Publication Ready Plots, (n.d.). https://rpkgs.datanovia.com/ggpubr/(accessed January 25, 2024).

[39] Add P-Values and Significance Levels to Ggplots - Articles - STHDA, 2017. http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/76-add-p-values-and-significance-levels-to-ggplots/. (Accessed 25 January 2024).

[40] N. Jones, J.F. Bohnsack, S. Takahashi, K.A. Oliver, M.-S. Chan, F. Kunst, P. Glaser, C. Rusniok, D.W.M. Crook, R.M. Harding, N. Bisharat, B.G. Spratt, Multilocus sequence typing system for group B streptococcus, J. Clin. Microbiol. 41 (2003) 2530–2536, https://doi.org/10.1128/JCM.41.6.2530-2536.2003.

[41] C. Crestani, T.L. Forde, J. Bell, S.J. Lycett, L.M.A. Oliveira, T.C.A. Pinto, C.G. Cobo-Ángel, A. Ceballos-Márquez, N.N. Phuoc, W. Sirimanapong, S.L. Chen, D. Jamrozy, S.D. Bentley, M. Fontaine, R.N. Zadoks, Three Accessory Gene Clusters Drive Host-Adaptation in Group B Streptococcus, 2023, https://doi.org/10.1101/2023.08.10.552778, 2023.08.10.552778.

[42] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explanations: an Overview of Interpretability of Machine Learning, 2019. http://arxiv.org/abs/1806.00069. (Accessed 2 May 2023).

[43] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: a review of machine learning interpretability methods, Entropy 23 (2021) 18, https://doi.org/10.3390/e23010018.

[44] Y. Ning, M.E.H. Ong, B. Chakraborty, B.A. Goldstein, D.S.W. Ting, R. Vaughan, N. Liu, Shapley variable importance cloud for interpretable machine learning, Patterns (N Y) 3 (2022) 100452, https://doi.org/10.1016/j.patter.2022.100452.

[45] E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher, G. Groh, SHAP-based explanation methods: a review for NLP interpretability, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2022, pp. 4593–4603. Gyeongju, Republic of Korea, https://aclanthology.org/2022.coling-1.406. (Accessed 2 May 2023).

[46] C.K. Glidden, A.R. Murran, R.A. Silva, A.A. Castellanos, B.A. Han, E.A. Mordecai, Phylogenetic and biogeographical traits predict unrecognized hosts of zoonotic leishmaniasis, PLoS Neglected Trop. Dis. 17 (2023) e0010879, https://doi.org/10.1371/journal.pntd.0010879.

[47] K.E.L. Worsley-Tonks, L.E. Escobar, R. Biek, M. Castaneda-Guzman, M.E. Craft, D. G. Streicker, L.A. White, N.M. Fountain-Jones, Using host traits to predict reservoir host species of rabies virus, PLoS Neglected Trop. Dis. 14 (2020) e0008940, https://doi.org/10.1371/journal.pntd.0008940.

[48] F. Mock, A. Viehweger, E. Barth, M. Marz, VIDHOP, viral host prediction with Deep Learning, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btaa705 btaa705.

[49] L. Brierley, A. Fowler, Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through

machine learning, Genomics (2020), https://doi.org/10.1101/2020.11.02.350439.
[50] M. Zhang, L. Yang, J. Ren, N.A. Ahlgren, J.A. Fuhrman, F. Sun, Prediction of virus-host infectious association by supervised learning methods, BMC Bioinf. 18 (2017) 60, https://doi.org/10.1186/s12859-017-1473-7.
[51] S. Roux, A.P. Camargo, F.H. Coutinho, S.M. Dabdoub, B.E. Dutilh, S. Nayfach, A. Tritt, iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria, PLoS Biol. 21 (2023) e3002083, https://doi.org/10.1371/journal.pbio.3002083.
[52] S.L. Chen, Genomic insights into the distribution and evolution of group B Streptococcus, Front. Microbiol. 10 (2019). https://www.frontiersin.org/articles/10.3389/fmicb.2019.01447. (Accessed 13 June 2023).
[53] C.D. Loaiza, N. Duhan, M. Lister, R. Kaundal, In silico prediction of host-pathogen protein interactions in melioidosis pathogen Burkholderia pseudomallei and human reveals novel virulence factors and their targets, Briefings Bioinf. 22 (2021) bbz162, https://doi.org/10.1093/bib/bbz162.
[54] T. Bharucha, B. Gangadharan, A. Kumar, A.C. Myall, N. Ayhan, B. Pastorino, A. Chanthongthip, M. Vongsouvath, M. Mayxay, O. Sengvilaipaseuth, O. Phonemixay, S. Rattanavong, D.P. O'Brien, I. Vendrell, R. Fischer, B. Kessler, L. Turtle, X. De Lamballerie, A. Dubot-Pérès, P.N. Newton, N. Zitzmann, SEAe consortium, deep proteomics network and machine learning analysis of human cerebrospinal fluid in Japanese encephalitis virus infection, J. Proteome Res. 22 (2023) 1614–1629, https://doi.org/10.1021/acs.jproteome.2c00563.
[55] C.A.G. Leal, G.A. Queiroz, F.L. Pereira, G.C. Tavares, H.C.P. Figueiredo, Streptococcus agalactiae sequence type 283 in farmed fish, Brazil, Emerg. Infect. Dis. 25 (2019) 776–779, https://doi.org/10.3201/eid2504.180543.