antibiotics

MDPI

Article

Source Attribution of Antibiotic Resistance Genes in Estuarine Aquaculture: A Machine Learning Approach

Helena Sofia Salgueiro [1], Ana Cristina Ferreira [2,3], Ana Sofia Ribeiro Duarte [4,*] and Ana Botelho [2,*]

[1] Faculty of Veterinary Medicine, University of Lisbon, 1300-477 Lisbon, Portugal; helenasofiasalgueiro@gmail.com
[2] National Institute for Agrarian and Veterinary Research (INIAV IP), Av. da República, Quinta do Marquês, 2780-157 Oeiras, Portugal; cristina.ferreira@iniav.pt
[3] BioISI—Instituto de Biosistemas e Ciências Integrativas, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal
[4] National Food Institute, Technical University of Denmark, Kemitorvet 204, 2800 Kongens Lyngby, Denmark
* Correspondence: asrd@food.dtu.dk (A.S.R.D.); ana.botelho@iniav.pt (A.B.)

Abstract: Aquaculture located in urban river estuaries, where other anthropogenic activities may occur, has an impact on and may be affected by the environment where they are inserted, namely by the exchange of antimicrobial resistance genes. The latter may ultimately, through the food chain, represent a source of resistance genes to the human resistome. In an exploratory study of the presence of resistance genes in aquaculture sediments located in urban river estuaries, two machine learning models were applied to predict the source of 34 resistome observations in the aquaculture sediments of oysters and gilt-head sea bream, located in the estuaries of the Sado and Lima Rivers and in the Aveiro Lagoon, as well as in the sediments of the Tejo River estuary, where Japanese clams and mussels are collected. The first model included all 34 resistomes, amounting to 53 different antimicrobial resistance genes used as source predictors. The most important antimicrobial genes for source attribution were tetracycline resistance genes *tet*(*51*) and *tet*(*L*); aminoglycoside resistance gene *aadA6*; beta-lactam resistance gene *blaBRO-2*; and amphenicol resistance gene *cmx_1*. The second model included only oyster sediment resistomes, amounting to 30 antimicrobial resistance genes as predictors. The most important antimicrobial genes for source attribution were the aminoglycoside resistance gene *aadA6*, followed by the tetracycline genes *tet*(*L*) and *tet*(*33*). This exploratory study provides the first information about antimicrobial resistance genes in intensive and semi-intensive aquaculture in Portugal, helping to recognize the importance of environmental control to maintain the integrity and the sustainability of aquaculture farms.

Keywords: aquaculture sediments; antimicrobial resistance; resistomes; source attribution; machine learning

Citation: Salgueiro, H.S.; Ferreira, A.C.; Duarte, A.S.R.; Botelho, A. Source Attribution of Antibiotic Resistance Genes in Estuarine Aquaculture: A Machine Learning Approach. *Antibiotics* **2024**, *13*, 107. https://doi.org/10.3390/antibiotics13010107

Academic Editor: Marc Maresca

Received: 29 November 2023
Revised: 12 January 2024
Accepted: 19 January 2024
Published: 22 January 2024

1. Introduction

Antibiotic residues can accumulate in the environment due to several anthropogenic activities, leading to selective pressure on bacteria from the environmental microbiome [1,2]. The occurrence of antibiotic resistance in the environment is a major concern due to the spread of resistant bacteria and resistance genes and the associated human health risks [3]. In fact, several international efforts are in place to incentivize the monitoring of antimicrobial resistance in the environment, e.g., the regular monitoring of antimicrobial resistance in the outlets of urban wastewater treatment plants in all agglomerations above 100,000 persons will become mandatory in the EU [4]. The presence of antimicrobial resistance genes (ARG) has been described in bacteria from several environments, namely in marine and aquaculture sediments and soil [5]. A recent meta-analysis study of 460 published articles revealed that the aquaculture sector is a reservoir of resistance to

antibiotics with therapeutic potential [1], and other recent scientific reviews [6,7] highlight the importance of the aquatic environment in the transmission of antimicrobial resistance.

Most of the studies about the spread of antibiotic resistance have been focused on acquired resistance [8]. However, in nature—specifically, in water and soil—antibiotic resistance is frequent and mostly intrinsic [9]. This is relevant in terms of antibiotic resistance ecology as it may create charity mechanisms that favor the acquisition of resistance by some community members [10], and some of the genes that are intrinsic in some species may become acquired in others [11]. Shotgun metagenomics sequencing approaches have contributed to the analysis of the span of antimicrobial resistance genes (ARGs) in a sample, irrespective of their intrinsic or acquired character, designated as the resistome [12].

The development and critical systematic assessment of genomic-based source attribution models of antimicrobial resistance (AMR) determinants enable the investigation of resistance gene transmission between several habitats and hosts. The analysis of genetic determinants of resistance derived from metagenomics sequencing, together with the appropriate epidemiological data, are valuable to determine the origin of AMR contamination in aquatic environments [13] and to predict possible contamination from, e.g., different animal reservoirs [14].

Aquaculture located in river estuaries, near industrial sites, farms and urban activities, is subjected to diverse anthropogenic influences and the likelihood of contamination with antimicrobial resistance is therefore high [15]. Sulfonamide resistance genes *sul1* and *sul2*, trimethoprim resistance gene *dfrA1* and class 1 integron *intI1* have persisted for six years in aquaculture sediments in Finland [16] and also in South Korea, China and Japan [17–19].

Most Portuguese aquaculture facilities focus on the production of marine fish and bivalve molluscs and operate essentially in estuaries and coastal lagoons, in intensive or semi-intensive systems. This is the case for the semi-intensive production of clams and oysters, which corresponded to over three quarters of the total mollusc production in Portugal in 2018 [20]. Oyster aquaculture has been growing in recent years in Portugal, and only a few studies have been performed to assess its impact on the surrounding ecosystem. A recent study has focused on the selection of antibiotic resistance by metals in riverine bacterial communities in Portugal [21].

In the comprehensive field of global health, machine learning tools have broadly intervened in morbidity and mortality risk assessment and the prediction of certain diseases' progression [22,23], infectious disease surveillance [24–26] and the improvement of health policy and planning [27,28]. For this reason, the interest in applying these tools to antimicrobial resistance genomic data has intensified over the past few years, not only reflecting the exponential increase in genomic AMR data available but also the increasing global awareness of the public health threat posed by AMR. In the current study, the machine learning random forest algorithm was applied to predict the relative attribution of antimicrobial resistance genes (ARGs) to different aquatic environments, based on 34 resistomes of sediments of aquaculture and of estuarine sediments where bivalves are captured.

2. Results

2.1. RF1—Source Attribution to Aquatic Environment Using Resistomes of Mussel, Gilt-Head Sea Bream and Oyster Aquaculture Sediments

After fitting the first random forest model with the train set, we obtained its performance metrics. The RF1 training model selected nine variables to use randomly at each split of the trees (mtry) and reached accuracy of 0.99. The confusion matrix of the Out-of-Bag (OOB) predictions showed that the model classified all samples correctly, resulting in an OOB error of 0%.

Similarly to training, a summary of the model's performance was obtained, after making predictions with the 25% hold-out set. The model incorrectly classified one sample, mistaking the Aveiro Lagoon for the Sado River, reaching an accuracy value of 0.95. The Mathews Correlation Coefficient (MCC) reached a value of 0.93.

Overall, the performance results demonstrated that the model was capable of correctly attributing different classes of aquatic environment to observations based on the abundance levels of various AMR determinants.

The built-in variable importance measure of RF resulted in the ranking of the twenty most important variables for the classification of RF1, which can be seen in Figure 1. Only six of these twenty variables were demonstrated to have high importance (importance value above 50). We focused our descriptive analysis of gene relative abundance on these six antimicrobial gene clusters, which included the tetracycline resistance genes *tet*(*51*), *tet*(*L*)*_1* and *tet*(*L*)*_4*; the aminoglycoside resistance gene *aadA6*; the beta-lactam resistance gene *blaBRO-2*; and the amphenicol resistance gene *cmx_1*. Among these genes, *tet*(*51*) reached an importance level of 100 and *aadA6* an importance level of 99, thus contributing together to the most prediction accuracy in RF1.

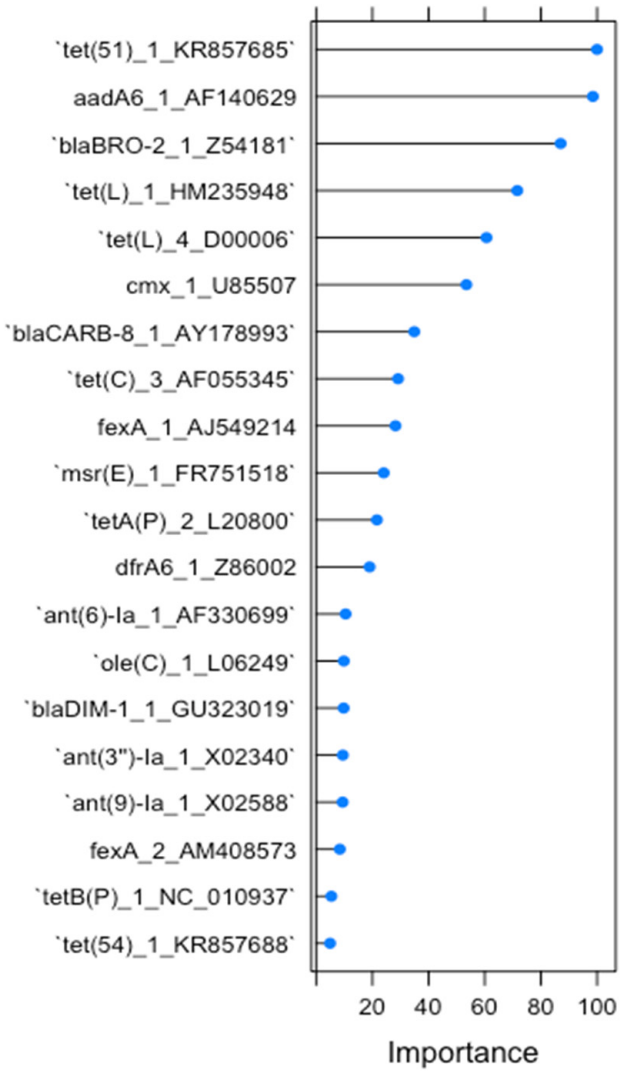

Figure 1. The 20 most important AMR determinants for classification in RF1.

The distribution of the relative abundance of the twenty most important variables in RF1 considering the environment of the aquaculture of mussels, gilt-head sea bream and oysters (Figure 2) demonstrated that most genes were present in only one aquatic environment.

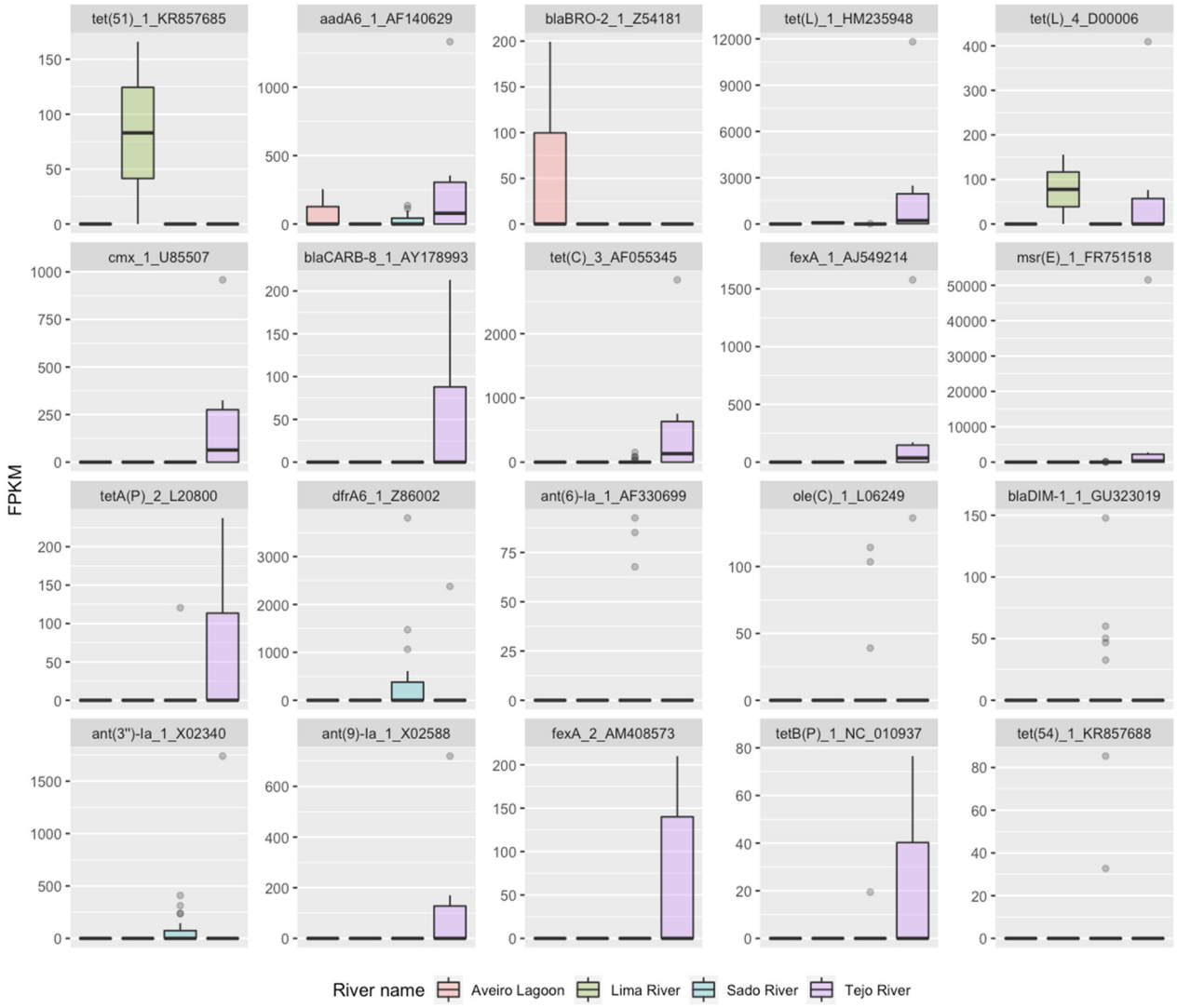

Figure 2. Distribution of abundance in FPKM values of the 20 most important AMR determinants for classification of aquatic environment in RF1.

Among the six genes with the highest importance, *tet*(*51*) was only present in the Lima estuary (oyster production), *blaBRO-2* was only present in the Aveiro Lagoon (oyster production), and *cmx_1* was only present in the Tejo estuary (clam and mussel collection), while *tet*(*L*) was present in two environments (Lima and Tejo estuaries), and *aadA6* was present in three (Aveiro Lagoon and Sado and Tejo estuaries). Ten out of the twenty genes with the top importance in RF1 were detected mostly or exclusively in the Tejo River and six in the Sado estuary (gilt-head bream and oyster aquaculture). In the Aveiro Lagoon and Lima estuary, only two out of the twenty genes were detected in each, *aadA6* and *blaBRO-2* in the first case and *tet*(*51*) and *tet*(*L*) in the latter.

2.2. RF2—Source Attribution to Aquatic Environment Using Resistomes of Oyster Aquaculture Sediments

The RF2 model training performance showed accuracy of 0.83. The model selected four random variables at each split of the trees. The confusion matrix of the OOB predictions demonstrated that the model incorrectly classified seven of the fifty-seven observations, resulting in an OOB error of 15.56%. Six of the misclassifications were between the Lima and Sado estuaries, and another between the Lima estuary and Aveiro Lagoon.

Regarding prediction performance with the 25% hold-out set, the RF2 model had accuracy of 0.83 and an MCC of 0.77. It incorrectly predicted two resistomes from the Sado estuary, classifying one as the Aveiro Lagoon and the other as the Lima estuary.

Compared to RF1, the RF2 performance results were not as positive, mostly because the MCC was significantly lower, indicating that the RF2 model might not be as reliable as RF1.

As explained previously for RF1, we were able to assess the twenty most important variables for RF2, shown in Figure 3. For RF2, only three antimicrobial resistance genes had importance above 50, including the aminoglycoside gene *aadA6* and the tetracycline genes *tet*(*L*) and *tet*(*33*). Gene *aadA6* reached an importance value of 100, being the most important gene for classification, while the following two genes, *tet*(*L*) and *tet*(*33*), had an importance level of 67 and 61, respectively.

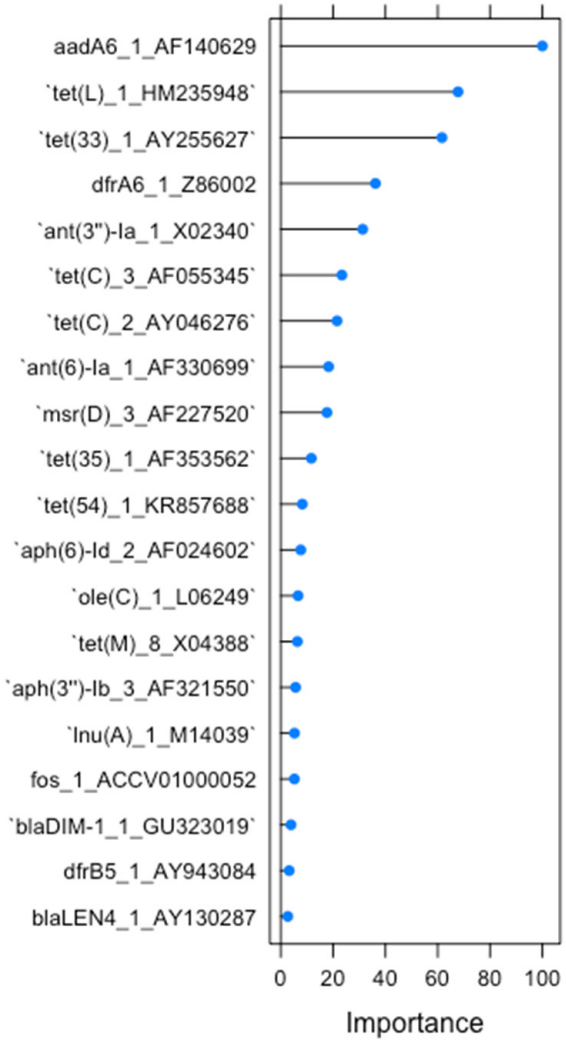

Figure 3. The 20 most important AMR determinants for classification in RF2.

We inspected the distribution of abundance of the twenty most important variables in RF2 (Figure 4). The three genes with the highest importance were the ones present in more than one environment. Genes *aadA6* and *tet*(*33*) were detected in the Aveiro Lagoon and Sado estuary and *tet*(*L*) was detected in the Sado and Lima estuaries. The remaining seventeen of the twenty most important resistance genes in RF2 were only detected in the Sado estuary, mostly sporadically, with a few exceptions that were often detected, including *dfrA6*, *ant*(*3″*)-*Ia*, *tet*(*C*) and *tet*(*35*). Among all twenty variables, only *tet*(*L*) was present in the Lima estuary, and only *aadA6* and *tet*(*33*) were present in the Aveiro Lagoon.

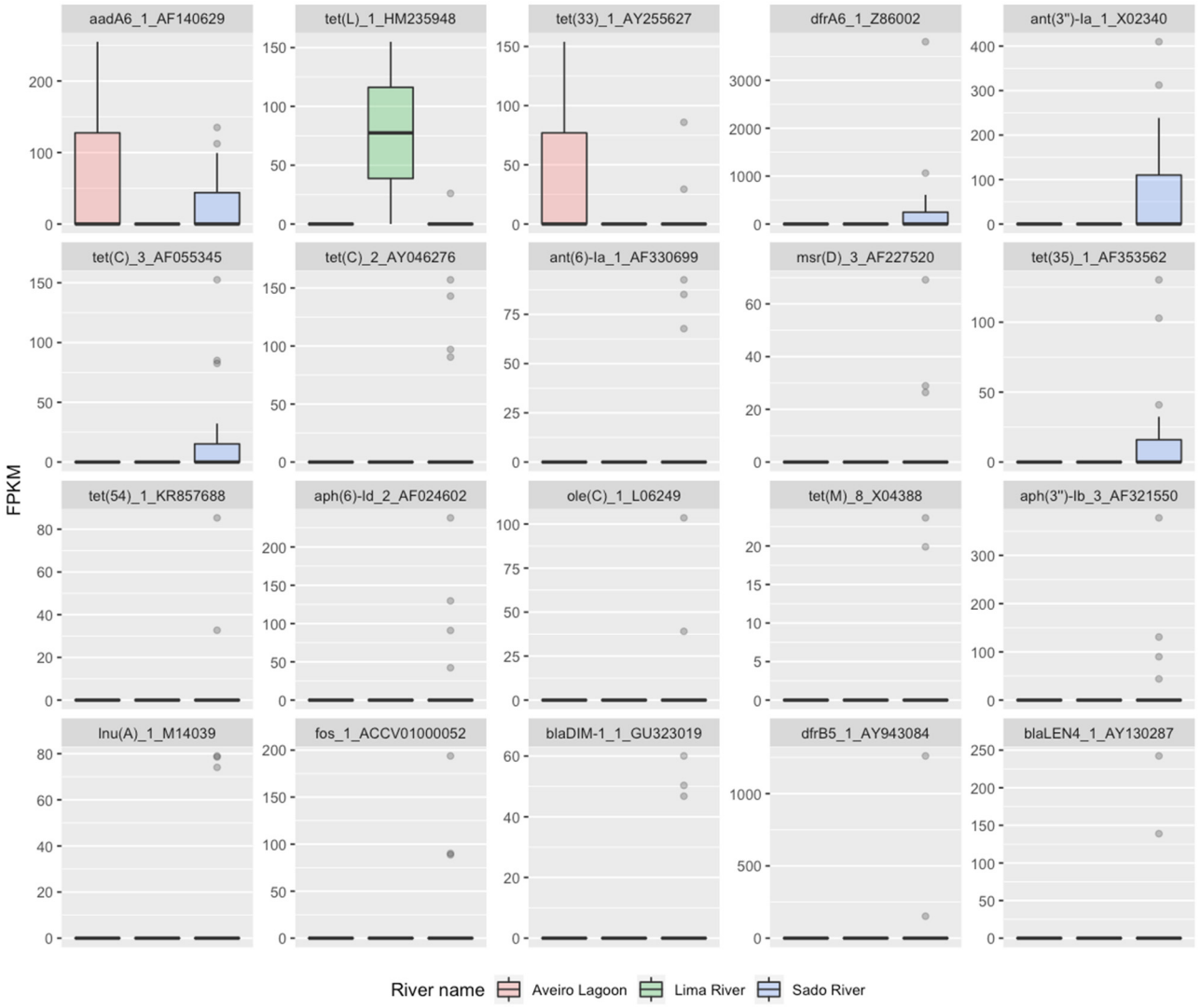

Figure 4. Distribution of abundance in FPKM values of the 20 most important genes for classification of aquatic environment in RF2.

3. Discussion

The health sector has increasingly benefited from the function of machine learning algorithms over the years, particularly for the study of AMR, due to the facilitated access to AMR genomic datasets [29]. In this exploratory study, we applied a random forest algorithm to a metagenome-based AMR dataset (resistome) with the purpose of linking the sources of AMR genes to different aquaculture systems located in the estuaries of rivers Sado, Tejo and Lima and in the Aveiro Lagoon.

Random forest has been demonstrated to be one of the most accurate classification algorithms for genomic data analysis, being applied in an increasing number of studies [30,31]. Despite being known for its high accuracy and ability to learn from extensive data, RF models have some limitations, which can lead, in specific scenarios, to suboptimal performance. The main limitation to the use of RF is its tendency to overfit, which is a critical problem when training machine learning models. Overfitting occurs when a model learns the training data too well, in such a way that it memorizes its noise and idiosyncrasies, instead of learning the underlying patterns. There are many causes of overfitting, the most common ones being an insufficient sample size and incorrect model tuning. A model too tuned for the training set (trees excessively deep, or too many trees) will lead to a complex and specific model for these samples only [32]. The present dataset is restricted to only 34 observations, which is a small sample size considering the purpose of fitting an RF

model. A small number of observations limits the learning content during the training process, which implies that the model will not be able to make accurate predictions with new observations; thus, it cannot be generalized to new datasets [30,31]. Moreover, these data are exceedingly unbalanced, both in terms of river location and aquaculture species. From the 34 observations, 70% were from oysters and 23 came from Rio Sado, 6 from Rio Tejo and only 2 from Rio Lima and 3 from the Aveiro Lagoon. Unbalanced data is one of the greatest challenges when it comes to RF models. Although we upsampled the data to guarantee the proportion of the same class samples, this was done by replicating the minority samples, and some studies suggest that oversampling very unbalanced data can also lead to overfitting, due to the similarity between the same class samples and low diversity of patterns [29].

Here, we applied two different models, RF1 and RF2, which differed in the number of aquaculture species included, and consequently in the number of samples.

Comparing both models' performance, RF1 demonstrated superior results compared to RF2. The high accuracy of the model RF1, considering the limited sample size and its unbalanced nature, most likely resulted from overfitting. In order to produce a reliable model, the number of samples collected from each river should be significantly higher and the distribution of samples among the different habitats should be balanced.

In this study, we assessed the 20 most important antimicrobial resistance genes for the model predictions and their abundance among river sites. For both models, the abundance of these genes was barely distributed among all rivers, showing most of the genes only present in the Sado estuary. This may be related to the fact that there is a greater number of samples from this river.

However, the genes with a higher level of importance (*tet*(*51*), *aadA6*, *blaBRO-2*, *tet*(*L*)*_1*, *tet*(*L*)*_4* and *cmx_1* for RF1; *aadA6*, *tet*(*L*)*_1* and *tet*(*33*) for RF2 appear to be those present in a wider variety of estuaries. This indicates that the model learned from the diversity of patterns and based its predictions on the different abundance of genes between aquatic environments.

The acquired genes *tet*(*L*) and *tet*(*33*) confer resistance to tetracycline and code for energy-dependent membrane-associated proteins that export tetracycline out of the cell [33]. Tetracyclines have become one of the most widely used classes of antibiotics in agriculture and aquaculture, due to their broad antimicrobial spectrum, oral availability and low cost. Extensive use over the past seven decades has selected for the expansion of tetracycline resistance in environmental microorganisms [34,35]. As a result of extensive anthropogenic use, tetracycline resistance is now widespread and has frequently been found in fresh water aquaculture [36–38].

No discriminatory AMR pattern was observed within the three main locations of oyster aquaculture (Sado and Lima estuaries and Aveiro lagoon), located several kms apart. Based on model RF2, a resistance gene fingerprint of Portuguese oyster estuarine aquaculture could include tetracycline (*tet*(*L*) and *tet*(*33*)) and aminoglycoside (*aadA6*) genes. Previously, Silva and colleagues [15] also reported resistance to tetracycline classes, applying a comparative genomics approach to the same metagenome samples.

Most of the specimens of these oyster aquaculture systems are for export and are controlled in terms of temperature, pH and metal ions. Thus, unless a sporadic event of wastewater treatment plants (WWTPs) or fabric discharge occurs, the resistome of the sediments in these production sites is expected to be somewhat constant. The regular monitoring of AMR patterns in the precise surrounding environments of such aquaculture could help to detect and trace back contamination events in time and space.

For example, in Sado's estuary, where most of the samples were collected, two regions could be identified: one surrounded by industrial activity (namely paper factories; tomato, milk and fertilizer producers; and two wastewater facilities) and another surrounded by small pig farms. In this second region, residues of enrofloxacin, a fluoroquinolone antibiotic used for the treatment or prophylaxis in pigs [39], were detected in the flesh of the gilt-head

bream from the aquaculture (results not shown), suggesting the occurrence of spillover contamination from the pig production environment to the aquaculture environment.

Furthermore, considering model RF1 and its 20 most important AMR genes, Tejo's estuary presented eleven of these genes, which is remarkable when compared with the other environments, where there was a predominance of only one gene. This distinct resistome might be explained by the non-controlled environment in an area with strong anthropogenic influence, where bivalves are collected by fishermen in a non-licensed place.

At first glance, RF2 seems to have a less confusing dataset for the model, as it predicts estuaries using the patterns of just one species, oysters, distributed among three estuaries. However, the results demonstrate that, even though the number of samples from mussels and gilt-head sea bream is small and they are present in only one estuary each, they help to increase the overall performance. We interpret that the presence of these samples increases the model performance because, since mussels and bream are only present in one estuary, they possibly affect the resistomes of the different aquatic environments considered in the model, thus increasing the sensitivity in prediction.

This study demonstrated that, despite the limitations in terms of sample size and balance, the resistome found in the sediment of aquaculture environments is possibly influenced by the species under production and probably consequently the different production practices. The differences encountered in the resistomes were sufficient to inform a random forest model that could successfully predict the aquatic environment of origin of a sediment sample. It also highlighted a clear difference in resistome composition between controlled aquaculture production sites and non-controlled environments where the unlicensed collection of bivalves occurs. In order to improve the model's validity and exclude possible confounding effects, larger datasets and relevant epidemiological data (e.g., proximity to farms or wastewater treatment plants), respectively, are needed in the future.

4. Materials and Methods

4.1. Sample Collection and DNA Extraction

Portuguese oyster aquaculture in the Sado and Lima estuaries and in the Aveiro Lagoon and gilt-head bream aquaculture in the Sado estuary were selected due to their importance and volume of production. In the Tejo estuary, a large water body surrounded by seven municipalities and industries with effluent discharges, Japanese clams and mussels are captured, and, due to this wild anthropogenic activity, this region was also selected.

Between November 2018 and July 2019, 24 samples of oyster (*Crassostrea angulata*) aquaculture sediments were collected in three estuarine regions, near urban centers: the Lima estuary, near Viana do Castelo city in the north of Portugal (two samples); the Aveiro Lagoon, near Aveiro city in the center (three samples); and the Sado estuary, near Setúbal city in the south (19 samples). Four sediments of gilt-head bream (*Sparus aurata*) aquaculture were also collected in Sado's estuary. In the Tejo River estuary, Japanese clams (*Ruditapes philippinarum*) and mussels (*Mytilus* spp.) are freely collected for human consumption, and, from there, six sediments were collected. All 34 sediment samples were stored at -20 °C until processing for total DNA extraction using a Qiagen PowerSoil Pro extraction kit, ID 47016 (Germany), according to the manufacturer's instructions. The DNA concentration and quality were determined using a Qubit 4.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA 02451,USA). High-quality DNA samples ($\geq$500 ng high molecular weight in a concentration of $\geq$20 ng μL^{-1}) in at least a 25 µL volume were used for library construction.

4.2. Metagenomic Sequencing, Reference Databases and Bioinformatic Analysis

Metagenomic sequencing was performed at CeGaT (GmbH Tuebingen, Germany) with the NovaSeq 6000 platform (Illumina, Inc., San Diego, CA 92122 USA), as well as whole genome sequencing, followed by the preparation of the sequencing libraries (Illumina DNA (M) Tagmentation Library Prep kit), with a read length of 2 $\times$ 100 bp, resulting in an output of about 20 billion bases (10 M clusters) per sample. Demultiplexing of the

sequencing reads was performed with Illumina bcl2fastq (2.20) and adapters were trimmed with Skewer (version 0.2.2). The quality of the FASTQ files was analyzed with FastQC (version 0.11.5-cegat) and the data were delivered as trimmed FASTQ files.

4.3. Machine-Learning Based Source-Attribution

All analyses conducted in this study were performed in R Studio v.1.3.1093, and the full study design is presented in Figure 5.

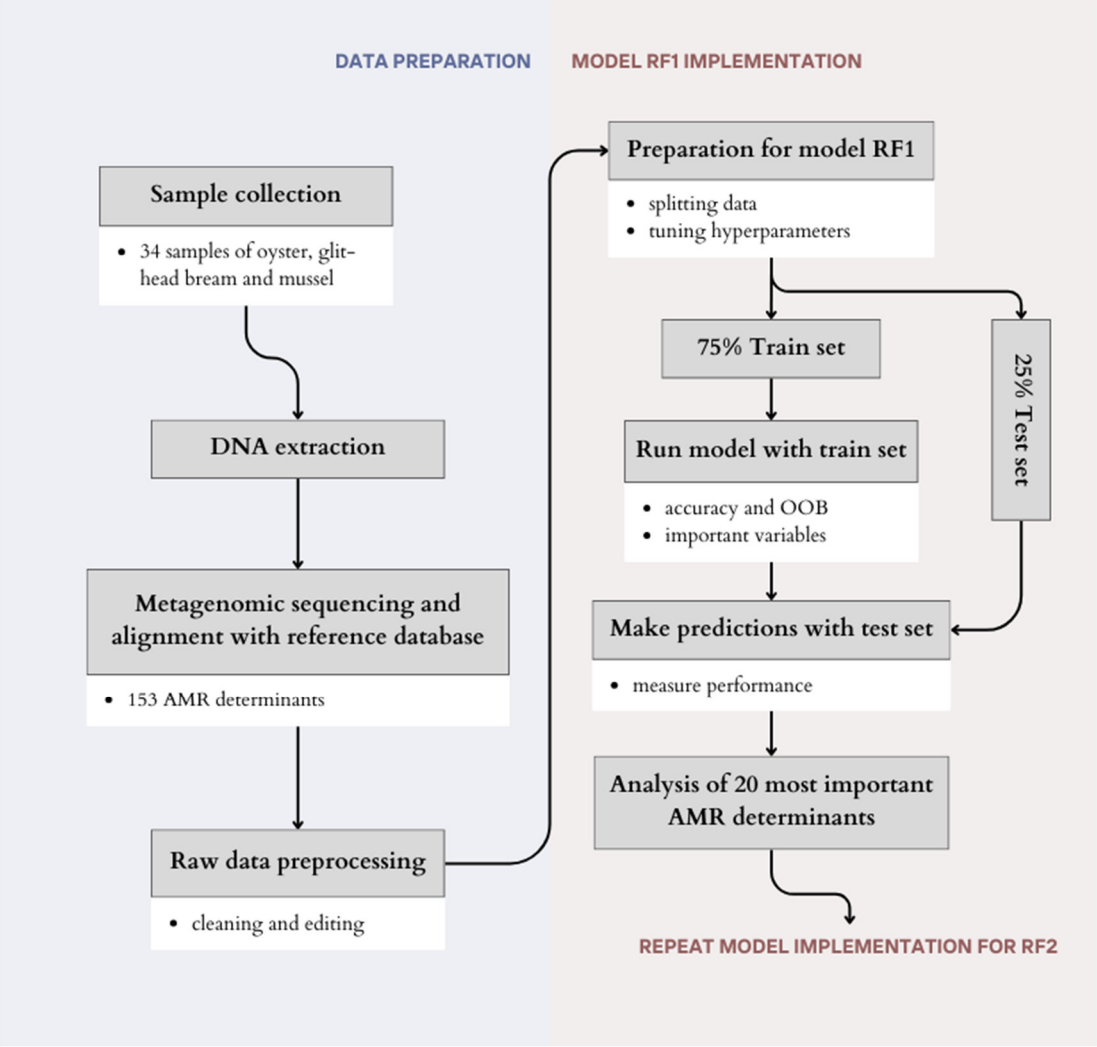

Figure 5. Study design.

The supervised machine learning algorithm random forest (RF) was applied to predict the class of a resistome observation regarding the aquaculture environment. Random forest is a tree-based supervised machine learning algorithm that is highly adaptive and is able to account for correlations and interactions among explanatory variables, also called features [40]. RF can be a collection of hundreds or thousands of decision trees. Each tree uses a bootstrap sample of the original data, and binary splits recursively partition the tree based on the most popular classification, pushing the samples from a parent node to its two daughter nodes, so that the homogeneity in the daughter nodes is improved [30]. The model uses a random number of features in each split to increase the accuracy and randomization. When this collection of trees is generated, a final vote is cast based on the classification of every tree [41]. Prediction for each observation is based on the proportion of votes given to each class across all trees, in the form of relative probabilities, from which the model produces a final classification based on the most likely class—the "crisp class" [42].

Two models were applied in this study, named RF1 and RF2, using the R package *caret* v.6.0-93 [43]. The two models used the same data; however, RF1 included all samples and RF2 included only oyster samples.

In terms of data preprocessing, all zero and near-zero variance features were removed, and highly correlated features were assessed and eliminated using the *findCorrelation* function of *caret* [43]. This function creates a correlation matrix and returns the pairwise correlations that are above a given cutoff and the features to be removed in order to reduce the level of multicorrelation in the data. A balanced proportion of samples of different classes was assured by upsampling the data. The approach of upsampling with replacement was used, where all original data are left intact and additional samples are added to the minority classes, with replacement [44].

In both models, tuning of the model's hyperparameters was performed, which consisted of (a) establishing 200 as the number of decision trees to be generated (*ntree*), (b) selecting 10-fold cross-validation repeated 5 times as a resampling method and (c) defining an interval of *mtry* values (number of variables used at each tree split). Each model then selected the *mtry* value that provided the best performance.

Both datasets were randomly split into a train set and a hold-out set, corresponding to 75% and 25% of the original data, respectively. Each model was trained with the 75% train set and using the *train* function of R package *caret* v6.0-93 [43], after which the algorithm presented the final model's characteristics, such as the *mtry* and model performance assessed based on accuracy and Out-of-Bag (OOB) error.

In RF, accuracy is defined as the number of correct classifications divided by the number of samples. The built-in accuracy metric of the *train* function ranges from 0 to 1 and is calculated using the k-fold cross-validation method, where the model is trained on k-1 folds and uses the remaining fold to test predictions. This process is repeated n times, after which the accuracy is then calculated as the average proportion of correctly classified instances over all folds [45].

The built-in OOB error in RF is described as the fraction of incorrect classifications over the number of out-of-bag samples. When each bootstrap sample is selected from the training data, the observations that are left out are called out-of-bag samples, which are extremely useful to estimate the generalization error and variable importance. Each OOB sample is passed down the tree to produce an estimated prediction error for the sample [46].

The final model was fit to the 25% hold-set to make predictions, using the *predict* function of the caret package. A confusion matrix was composed based on crisp-class classifications, and the prediction performance was assessed using the accuracy and Mathews Correlation Coefficient (MCC). The MCC is a statistical metric used for model evaluation, calculated based on the occurrence of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) [47]. This metric ranges from -1 to 1, where 1 indicates perfect agreement between the prediction and observation, -1 indicates total disagreement and 0 is expected for a prediction no better than random. The MCC was used because it has been shown to be more reliable than other performance metrics, such as the RF accuracy values [48].

After evaluating the model performance, the twenty most important variables for classification were analyzed. The RF importance measures help to understand the significance of each variable in the model [49]. For this task, the *varImp* function of the *caret* R package was used [43]. In each decision tree, every time a variable splits a node, the reduction in impurity (Gini impurity) in the two daughter nodes is calculated. The Gini impurity is a measure of the randomness in the split of the decision tree [50]. The *varImp* function calculates the importance score of each variable by averaging this reduction in impurity caused by the variable over all decision trees in the RF model [43]. The more a variable contributes to the reduction in impurity, the higher its importance score. The scores are then scaled such that they range from 0 to 100, to make it easier to compare the importance scores of different variables.

The distribution of abundance of the top twenty genes with the highest importance was plotted for each class of observations, to visualize differences in abundance between classes. Details of each generated model are described below.

4.3.1. RF1—Source Attribution to Aquatic Environment Using Mussel, Gilt-Head Bream and Oyster Resistomes

The first random forest included all data available. These data consisted of 34 resistome observations from oysters, mussels and gilt-head bream, distributed among 4 aquatic environments situated in Portugal (Sado, Tejo and Lima estuaries, as well as Aveiro Lagoon). Among all resistomes, 153 different antimicrobial resistance genes were detected. After upsampling and the removal of features with near-zero variance and highly correlated features, the data consisted of 92 observations, 23 of each environment and 53 features (antimicrobial resistance genes). After data splitting, the train set consisted of 52 observations and the hold-out set of 20 observations. The model was fit using an interval from 4 to 12 as the *mtry*, in order to include, with some margin, the square root of the number of features (rule of thumb of *mtry* values).

4.3.2. RF2—Source Attribution to Aquatic Environment Using Oyster Resistomes

To exclude possible species-related resistome variations, a single species was selected to run a second random forest (RF2). Due to the larger representation of different aquatic environments among oyster samples, a second model was performed including only observations of oyster resistomes. Oyster samples were collected in 3 of the 4 environments; thus, RF2 had 3 output classes (Sado and Lima estuaries and Aveiro Lagoon). Originally with 24 observations, after upsampling and the removal of predictor features with zero variance, the data consisted of 57 observations, distributed among 3 aquatic environments, and 30 predictor antimicrobial resistance genes. After splitting, the train set consisted of 45 observations and the hold-out set of 12 observations. Maintaining the same criteria used in RF1, the interval of values used as *mtry* ranged from 2 to 10.

Author Contributions: Conceptualization, A.B. and A.S.R.D.; methodology, A.S.R.D., A.C.F. and H.S.S.; data curation, A.B.; writing—original draft preparation, A.B. and A.C.F.; writing—review and editing, A.B., A.S.R.D., H.S.S. and A.C.F.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundação para a Ciência e a Tecnologia, IP (FCT)/MCTES through national funds (PIDDAC) and co-funding by the European Regional Development Fund (FEDER) of the European Union, through the Lisbon Regional Operational Program and the Competitiveness and Internationalization Operational Program for Portugal 2020 in the scope of project "AquaRAM: Antimicrobial Resistance determinants in aquaculture environments" (references ALG-01-0145-FEDER-028824 and PTDC/BIA-MIC/28824/2017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://www.researchgate.net/publication/365620017_Database_of_Metgenomes_of_Sediments_from_Estuarine_Aquaculture_Farms_in_Portugal-AquaRAM_Project_Collection accessed on 11 November 2022.

Acknowledgments: Thanks are due to Tine Hald for hosting a short-term mission (STM-2022), of which Ana Cristina Ferreira was a recipient, at DTU Food (Denmark) in the framework of the project OHEJP-DISCOVER-WP3. Ana Luísa Maulvaud (IPMA), Andreia Freitas (INIAV), Filomena Soares (IPMA—EPPO), Leonor Orge (INIAV), Patricia Anacleto (IPMA), Sandra Cavaco Goncalves (INIAV) and Teresa Nogueira (INIAV) are acknowledged for their contributions in the collection and preparation of sediment samples, and Susana Lopes (CIBIO) for the support in DNA extraction from sediments.

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

References

1. Reverter, M.; Sarter, S.; Caruso, D.; Avarre, J.-C.; Combe, M.; Pepey, E.; Pouyaud, L.; Vega-Heredía, S.; de Verdal, H.; Gozlan, R.E. Aquaculture at the crossroads of global warming and antimicrobial resistance. *Nat. Commun.* **2020**, *11*, 1870. [CrossRef] [PubMed]
2. Watts, R.; Day, C.; Krzanowski, J.; Nutt, D.; Carhart-Harris, R. Patients' Accounts of Increased "Connectedness" and "Acceptance" After Psilocybin for Treatment-Resistant Depression. *J. Humanist. Psychol.* **2017**, *57*, 520–564. [CrossRef]
3. Koch, N.; Islam, N.F.; Sonowal, S.; Prasad, R.; Sarma, H. Environmental antibiotics and resistance genes as emerging contaminants: Methods of detection and bioremediation. *Curr. Res. Microb. Sci.* **2021**, *2*, 100027. [CrossRef] [PubMed]
4. European Comission. *Proposal for a Directive of the European Parliament and of the Council Concerning Urban Wastewater Treatment (Recast)*; COM/2022/0541; European Commission: Brussels, Belgium, 2022.
5. Helsens, N.; Calvez, S.; Prevost, H.; Bouju-Albert, A.; Maillet, A.; Rossero, A.; Hurtaud-Pessel, D.; Zagorec, M.; Magras, C. Antibiotic Resistance Genes and Bacterial Communities of Farmed Rainbow Trout Fillets (Oncorhynchus mykiss). *Front Microbiol.* **2020**, *11*, 590902. [CrossRef]
6. EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards); Koutsoumanis, K.; Allende, A.; Álvarez-Ordóñez, A.; Bolton, D.; Bover-Cid, S.; Chemaly, M.; Davies, R.; De Cesare, A.; Herman, L.; et al. Role played by the environment in the emergence and spread of antimicrobial resistance (AMR) through the food chain. *EFSA J.* **2021**, *19*, 188. [CrossRef]
7. Carvalho, I.T.; Santos, L. Antibiotics in the aquatic environments: A review of the European scenario. *Environ. Int.* **2016**, *94*, 736–757. [CrossRef] [PubMed]
8. Ventola, C.L. The Antibiotic Resistance Crisis: Part 1: Causes and threats. *Pharm. Ther.* **2015**, *40*, 277–283.
9. Jia, W.-L.; Song, C.; He, L.-Y.; Wang, B.; Gao, F.-Z.; Zhang, M.; Ying, G.-G. Antibiotics in soil and water: Occurrence, fate, and risk. *Curr. Opin. Environ. Sci. Health* **2023**, *32*, 100437. [CrossRef]
10. Bottery, M.J.; Pitchford, J.W.; Friman, V.-P. Ecology and evolution of antimicrobial resistance in bacterial communities. *ISME J.* **2020**, *15*, 939–948. [CrossRef]
11. Li, L.-G.; Yin, X.; Zhang, T. Tracking antibiotic resistance gene pollution from different sources using machine-learning classification. *Microbiome* **2018**, *6*, 93. [CrossRef]
12. de Abreu, V.A.C.; Perdigão, J.; Almeida, S. Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview. *Front. Genet.* **2021**, *11*, 575592. [CrossRef] [PubMed]
13. Gupta, S.; Arango-Argoty, G.; Zhang, L.; Pruden, A.; Vikesland, P. Identification of discriminatory antibiotic resistance genes among environmental resistomes using extremely randomized tree algorithm. *Microbiome* **2019**, *7*, 1–15. [CrossRef] [PubMed]
14. Duarte, A.S.R.; Röder, T.; Van Gompel, L.; Petersen, T.N.; Hansen, R.B.; Hansen, I.M.; Bossers, A.; Aarestrup, F.M.; Wagenaar, J.A.; Hald, T. Metagenomics-Based Approach to Source-Attribution of Antimicrobial Resistance Determinants—Identification of Reservoir Resistome Signatures. *Front. Microbiol.* **2021**, *11*, 601407. [CrossRef] [PubMed]
15. Silva, D.G.; Domingues, C.P.F.; Figueiredo, J.F.; Dionisio, F.; Botelho, A.; Nogueira, T. Estuarine Aquacultures at the Crossroads of Animal Production and Antibacterial Resistance: A Metagenomic Approach to the Resistome. *Biology* **2022**, *11*, 1681. [CrossRef] [PubMed]
16. Muziasari, W.I.; Pitkänen, L.K.; Sørum, H.; Stedtfeld, R.D.; Tiedje, J.M.; Virta, M. The Resistome of Farmed Fish Feces Contributes to the Enrichment of Antibiotic Resistance Genes in Sediments below Baltic Sea Fish Farms. *Front. Microbiol.* **2017**, *7*, 2137. [CrossRef] [PubMed]
17. Cabello, F.C.; Godfrey, H.P.; Tomova, A.; Ivanova, L.; Dölz, H.; Millanao, A.; Buschmann, A.H. Antimicrobial use in aquaculture re-examined: Its relevance to antimicrobial resistance and to animal and human health. *Environ. Microbiol.* **2013**, *15*, 1917–1942. [CrossRef] [PubMed]
18. Salgueiro, V.; Manageiro, V.; Bandarra, N.M.; Reis, L.; Ferreira, E.; Caniça, M. Bacterial Diversity and Antibiotic Susceptibility of *Sparus aurata* from Aquaculture. *Microorganisms* **2020**, *8*, 1343. [CrossRef]
19. Watts, J.E.M.; Schreier, H.J.; Lanska, L.; Hale, M.S. The Rising Tide of Antimicrobial Resistance in Aquaculture: Sources, Sinks and Solutions. *Mar. Drugs* **2017**, *15*, 158. [CrossRef]
20. Rocha, C.P.; Cabral, H.N.; Marques, J.C.; Gonçalves, A.M.M. A Global Overview of Aquaculture Food Production with a Focus on the Activity's Development in Transitional Systems—The Case Study of a South European Country (Portugal). *J. Mar. Sci. Eng.* **2022**, *10*, 417. [CrossRef]
21. Silva, I.; Tacão, M.; Henriques, I. Selection of antibiotic resistance by metals in a riverine bacterial community. *Chemosphere* **2021**, *263*, 127936. [CrossRef]
22. Phakhounthong, K.; Chaovalit, P.; Jittamala, P.; Blacksell, S.D.; Carter, M.J.; Turner, P.; Chheng, K.; Sona, S.; Kumar, V.; Day, N.P.J.; et al. Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: Application of classification tree analysis. *BMC Pediatr.* **2018**, *18*, 109. [CrossRef] [PubMed]
23. Johnston, I.G.; Hoffmann, T.; Greenbury, S.F.; Cominetti, O.; Jallow, M.; Kwiatkowski, D.; Barahona, M.; Jones, N.S.; Casals-Pascual, C. Precision identification of high-risk phenotypes and progression pathways in severe malaria without requiring longitudinal data. *NPJ Digit. Med.* **2019**, *2*, 63. [CrossRef] [PubMed]
24. Kabaria, C.W.; Molteni, F.; Mandike, R.; Chacky, F.; Noor, A.M.; Snow, R.W.; Linard, C. Mapping intra-urban malaria risk using high resolution satellite imagery: A case study of Dar es Salaam. *Int. J. Health Geogr.* **2016**, *15*, 1–12. [CrossRef] [PubMed]
25. Haddawy, P.; Hasan, A.I.; Kasantikul, R.; Lawpoolsri, S.; Sa-Angchai, P.; Kaewkungwal, J.; Singhasivanon, P. Spatiotemporal Bayesian networks for malaria prediction. *Artif. Intell. Med.* **2018**, *84*, 127–138. [CrossRef] [PubMed]

26. Clemente, L.; Lu, F.; Santillana, M. Improved Real-Time Influenza Surveillance: Using Internet Search Data in Eight Latin American Countries. *JMIR Public Health Surveill.* **2019**, *5*, e12214-58. [CrossRef] [PubMed]
27. Rosas, M.A.; Bezerra, A.F.B.; Duarte-Neto, P.J. Use of artificial neural networks in applying methodology for allocating health resources. *Public Health Prac.* **2013**, *47*, 128–136. [CrossRef]
28. Yousefi, M.; Yousefi, M.; Ferreira, R.P.M.; Kim, J.H.; Fogliatto, F.S. Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. *Artif. Intell. Med.* **2018**, *84*, 23–33. [CrossRef]
29. Elyan, E.; Hussain, A.; Sheikh, A.; Elmanama, A.A.; Vuttipittayamongkol, P.; Hijazi, K. Antimicrobial Resistance and Machine Learning: Challenges and Opportunities. *IEEE Access* **2022**, *10*, 31561–31577. [CrossRef]
30. Chen, X.; Ishwaran, H. Random forests for genomic data analysis. *Genomics* **2012**, *99*, 323–329. [CrossRef]
31. Luan, J.; Zhang, C.; Xu, B.; Xue, Y.; Ren, Y. The predictive performances of random forest models with limited sample size and different species traits. *Fish. Res.* **2020**, *227*, 105534. [CrossRef]
32. Ghojogh, B.; Crowley, M. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. Cornell University. *arXiv* **2019**, arXiv:1905.12787. [CrossRef]
33. Chopra, I.; Roberts, M.C. Tetracycline antibiotics: Mode of action, applications, molecular biology and epidemiology of bacterial resistance. *Microbiol. Mol. Biol. Rev.* **2001**, *65*, 232–260. [CrossRef] [PubMed]
34. Gasparrini, A.J.; Markley, J.L.; Kumar, H.; Wang, B.; Fang, L.; Irum, S.; Symister, C.T.; Wallace, M.; Burnham, C.-A.D.; Andleeb, S.; et al. Tetracycline-inactivating enzymes from environmental, human commensal, and pathogenic bacteria cause broad-spectrum tetracycline resistance. *Commun. Biol.* **2020**, *3*, 241. [CrossRef] [PubMed]
35. Ungemach, F.R.; Müller-Bahrdt, D.; Abraham, G. Guidelines for prudent use of antimicrobials and their implications on antibiotic usage in veterinary medicine. *Int. J. Med. Microbiol.* **2006**, *296* (Suppl. S41), 33–38. [CrossRef]
36. Gao, P.; Mao, D.; Luo, Y.; Wang, L.; Xu, B.; Xu, L. Occurrence of sulfonamide and tetracycline-resistant bacteria and resistance genes in aquaculture environment. *Water Res.* **2012**, *46*, 2355–2364. [CrossRef] [PubMed]
37. Xiong, W.; Sun, Y.; Zhang, T.; Ding, X.; Li, Y.; Wang, M.; Zeng, Z. Antibiotics, Antibiotic Resistance Genes, and Bacterial Community Composition in Fresh Water Aquaculture Environment in China. *Microb. Ecol.* **2015**, *70*, 425–432. [CrossRef] [PubMed]
38. Wang, Q.; Mao, C.; Lei, L.; Yan, B.; Yuan, J.; Guo, Y.; Li, T.; Xiong, X.; Cao, X.; Huang, J.; et al. Antibiotic resistance genes and their links with bacteria and environmental factors in three predominant freshwater aquaculture modes. *Ecotoxicol. Environ. Saf.* **2022**, *241*, 113832. [CrossRef] [PubMed]
39. Janssen, P.; Barton, G.; Kietzmann, M.; Meißner, J. Treatment of pigs with enrofloxacin via different oral dosage forms—environmental contaminations and resistance development of *Escherichia coli*. *J. Vet Sci.* **2022**, *23*, e23. [CrossRef]
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
41. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Ensemble Machine Learning*; Zhang, C., Ma, Y.Q., Eds.; Springer: Boston, MA, USA, 2012.
42. Biau, G.; Devroye, L.; Lugosi, G. Consistency of Random Forests and Other Averaging Classifiers. *J. Mach. Learn. Res.* **2008**, *9*, 2015–2033.
43. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
44. Ling, C.X.; Li, C. Data mining for direct marketing: Problems and solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD), New York, NY, USA, 27–31 August 1998; pp. 73–79.
45. Ljumovic, M.; Klar, M. Estimating expected error rates of random forest classifiers: A comparison of cross-validation and bootstrap. In Proceedings of the 4th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 14–18 June 2015; pp. 212–215.
46. Goldstein, B.A.; Hubbard, A.E.; Cutler, A.; Barcellos, L.F. An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genet.* **2010**, *11*, 49. [CrossRef]
47. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
48. Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.* **2018**, *174*, 33–44. [CrossRef]
49. Kuhn, M.; Johnson, K. Measuring Predictor Importance. In *Applied Predictive Modeling*, 1st ed.; Springer: New York, NY, USA, 2013; pp. 463–485. [CrossRef]
50. Yuan, Y.; Wu, L.; Zhang, X. Gini-Impurity Index Analysis. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 3154–3169. [CrossRef]