pathogens

MDPI

Article

A Machine Learning Model for Food Source Attribution of *Listeria monocytogenes*

Collins K. Tanui [1,2], Edmund O. Benefo [1], Shraddha Karanth [1] and Abani K. Pradhan [1,2,*]

[1] Department of Nutrition and Food Science, University of Maryland, College Park, MD 20742, USA; ctanui@umd.edu (C.K.T.); ebenefo@umd.edu (E.O.B.); skm@umd.edu (S.K.)
[2] Center for Food Safety and Security Systems, University of Maryland, College Park, MD 20742, USA
* Correspondence: akp@umd.edu

Abstract: Despite its low morbidity, listeriosis has a high mortality rate due to the severity of its clinical manifestations. The source of human listeriosis is often unclear. In this study, we investigate the ability of machine learning to predict the food source from which clinical *Listeria monocytogenes* isolates originated. Four machine learning classification algorithms were trained on core genome multilocus sequence typing data of 1212 *L. monocytogenes* isolates from various food sources. The average accuracies of random forest, support vector machine radial kernel, stochastic gradient boosting, and logit boost were found to be 0.72, 0.61, 0.7, and 0.73, respectively. Logit boost showed the best performance and was used in model testing on 154 *L. monocytogenes* clinical isolates. The model attributed 17.5 % of human clinical cases to dairy, 32.5% to fruits, 14.3% to leafy greens, 9.7% to meat, 4.6% to poultry, and 18.8% to vegetables. The final model also provided us with genetic features that were predictive of specific sources. Thus, this combination of genomic data and machine learning-based models can greatly enhance our ability to track *L. monocytogenes* from different food sources.

Keywords: *Listeria monocytogenes*; food source attribution; whole-genome sequencing; machine learning; predictive modeling

Citation: Tanui, C.K.; Benefo, E.O.; Karanth, S.; Pradhan, A.K. A Machine Learning Model for Food Source Attribution of *Listeria monocytogenes*. *Pathogens* **2022**, *11*, 691. https://doi.org/10.3390/pathogens11060691

Academic Editor: Patrick Njage

Received: 7 April 2022
Accepted: 10 June 2022
Published: 16 June 2022

1. Introduction

Foodborne illnesses affect approximately 48 million people in the United States every year, resulting in an estimated 128,000 hospitalizations and 3000 deaths [1]. About a fifth (approximately 9.4 million) of these can be attributed to known pathogens [2,3]. In most outbreak investigations, disease etiology is linked to individual foods, which enables public health authorities, regulatory agencies, and the food industry to identify potential points of contamination. Foodborne outbreak data can also be used to identify emerging food safety concerns and evaluate the effectiveness of foodborne illness prevention programs [4]. Foods of animal origin, fruits, and vegetables are usually implicated in most foodborne outbreaks [2,5,6]. Common pathogenic bacteria responsible for foodborne outbreaks include *Listeria monocytogenes*, *Campylobacter*, *Salmonella*, and Shiga toxin-producing *Escherichia coli*, among others [3,7].

L. monocytogenes causes serious illness only in a small percentage of healthy people. According to the United States Centers for Disease Control and Prevention (CDC), about 1600 people get listeriosis annually, and about 260 succumb to it [8]. Even though the number of listeriosis cases is lower than that of other foodborne illnesses, the disease burden of this pathogen is higher because of the serious nature of the disease when vulnerable groups are affected [8,9]. Listeriosis is ranked third among the causes of foodborne illness-associated deaths in the United States, causing nearly 19% of these deaths [10]. People who are at risk for listeriosis include pregnant women, the elderly, people with weakened immune systems, and newborns [10].

Food animals, particularly ruminants, can get infected with *L. monocytogenes*, making them potential zoonotic reservoirs of this pathogen [11,12]. Human infections are rarely related to exposure to infected animals or fomites from agricultural environments. However, animal-derived food products eaten raw or undercooked and refrigerated ready-to-eat (RTE) foods stored for long periods are known to cause listeriosis in humans [13,14].

Fresh produce is another food group that is gradually becoming a major route of human exposure to *L. monocytogenes* [15,16]. Unlike other foodborne pathogens, *L. monocytogenes* can thrive under alternative (i.e., non-ideal) conditions, such as low moisture, high salt concentration, and refrigeration temperature environments [17]. Since 2010, over 85 multistate outbreaks with confirmed etiology have been attributed to fresh produce in the United States [8]. Cross-contamination within the supply chain, improper storage temperatures during distribution, and improper food preparation practices are some of the frequently implicated contributors to these events.

Food source attribution is the process of estimating the most common food categories responsible for illnesses caused by specific pathogens [18,19]. Source attribution enables the identification of the relative contributions of different food sources to the occurrence of foodborne illnesses [20,21]. To achieve this, several sources of data are required including epidemiological, laboratory-, and outbreak-related data [22,23]. Unraveling the sources of foodborne illness is vital to identifying strategies to improve food safety along the entire food production and supply chain [19,24].

Multilocus sequence typing (MLST) [25,26] has been the preferred method for population genetic analyses, with the results usually corroborating epidemiological findings [26,27]. This molecular technique has been used to monitor changes in food microbial reservoirs, particularly those changes that arise as a result of interventions targeting the food chain and public health [26,28–30]. According to a prior study [26], core genome MLST (cgMLST) and allelic variations can be used to differentiate isolates and link them to food sources in source attribution studies. To decrease the prevalence of foodborne diseases and minimize microbial contamination in food, effective monitoring of the distribution and occurrence of foodborne pathogens is essential. It is worth noting that foodborne pathogens are resilient; this means that they can adapt genetically and phenotypically to the extreme conditions found in host and non-host systems, which allows them to survive and proliferate under these conditions [3,31,32]. These changes could be particularly informative towards identifying the basis of pathogen adaptation to, and survival and virulence in, host systems, as well as their response to safe food handling practices in the industry and by consumers. Therefore, a careful analysis of these changes could, in the long run, help develop methods and practices to reduce the risk of foodborne outbreaks.

In recent years, there has been a growing interest in analyzing genome sequencing data using artificial intelligence (AI), particularly machine learning (ML) [33]. Mechanistic model-based methods are aimed at formulating simplified mathematical models to explain various phenomena by carefully examining, analyzing, and identifying patterns in relevant data [34]. On the other hand, ML focuses on 'learning' from relevant patterns in data, and using this information to make predictions [35,36]. Basically, by exploring and identifying patterns in data, ML can be used in the classification, regression, or clustering of data to draw meaningful inferences from the same. Genome sequencing information, coupled with machine learning, has been used to predict the risk of listeriosis in humans [37], the host specificity of *S. enterica* and *E. coli* [38], and host disease severity based on *S. enterica* gene presence/absence [36,39], and in the source attribution of *S.* Typhimurium [26]. With the increase in usage of genome sequencing for exploratory and integrated surveillance activities, generating massive amounts of data, as well as standardization of data collection activities (providing us with useful metadata and other useful information), machine learning and big data analytical tools become the need of the hour to provide a better understanding and improvement of current knowledge in foodborne disease epidemiology.

This study aimed at developing a ML-based model for source attribution of human listeriosis by analyzing *L. monocytogenes* core genomes. The model was based on cgMLST

profiles from clinical *L. monocytogenes* isolates and isolates from dairy, fruits, leafy greens, meat, poultry, seafood, and vegetables.

2. Results

2.1. Predictive Model

We developed a supervised machine learning model to predict the possible source of human listeriosis cases based on allelic variations in *L. monocytogenes* isolates from foods. Of the 1748 *L. monocytogenes* core genes, 1012 genes were removed due to zero or near-zero variance (see Section 4.3.1), leaving 736 genes that were used in the model.

The performance of random forest, logit boost, stochastic gradient boosting, and support vector machine radial kernel models were compared using the average accuracies obtained from 10 iterations applying 10-fold cross-validation. All four models performed well with accuracies between 0.614 and 0.732, and Kappa values between 0.530 and 0.657 (Table 1).

Table 1. Models performance from 10 iterations of random forest, support vector machine radial kernel, stochastic gradient boosting, and logit boost models.

Models	Accuracy	95% CI	Kappa
Logit boost	0.732 [a]	0.665–0.760	0.654
Random forest	0.722 [a]	0.667–0.776	0.657
Stochastic gradient boosting	0.701 [a]	0.645–0.745	0.633
Support vector machine	0.614 [b]	0.569–0.671	0.530

Values under the Accuracy column with different superscripts are significantly different ($p < 0.05$).

The performance of logit boost (0.732), random forest (0.722), and stochastic gradient boosting (0.701) did not differ significantly from one another. However, these three models performed significantly better than support vector machine. Receiver operating characteristic (ROC) curves were generated for the different models. The areas under the curve (AUC) for logit boost, random forest, gradient boosting machine, and support vector machine were 0.865, 0.805, 0.822, and 0.820, respectively. Logit boost had the highest accuracy and AUC among the models considered and was selected for further analysis.This selection was also substantiated by the Kappa value for the logit boost model (0.654), suggesting a 'substantial' agreement between the observed and predicted classes [40] or a 'fair to good' agreement based on Fleiss's criteria [41].

Confusion matrix statistics of all train-test models are presented in Supplementary Table S1. Logit boost, the best performing model, had a specificity > 0.90 for all food sources, and sensitivity > 0.7 for most food sources, except leafy greens (0.548), meat (0.484), and poultry (0.447). The low sensitivity observed in leafy greens, meat, and poultry could be due to the smaller sample size in these categories as compared to the other categories. In the future, with the availability of an increased number of samples, especially in the less dominant classes, it may be possible to increase the sensitivity of the model for these classes. Other methods to potentially further improve the sensitivity of classifiers may include the use of resampling techniques and cost-sensitive learning approaches in future studies.

2.2. Source Attribution of Human Listeriosis Cases

We trained a new model using logit boost on the complete feature-reduced data set (Supplementary Table S2). This model predicted the probable food sources of each of the 154 clinical *L. monocytogenes* isolates. The model predicted that 32.5% of the clinical isolates originated from fruits, 18.8% from vegetables, and 17.5%, 14.3%, 9.7%, 4.6%, and 2.6% from dairy, leafy greens, meat, poultry, and seafood, respectively (Figure 1).

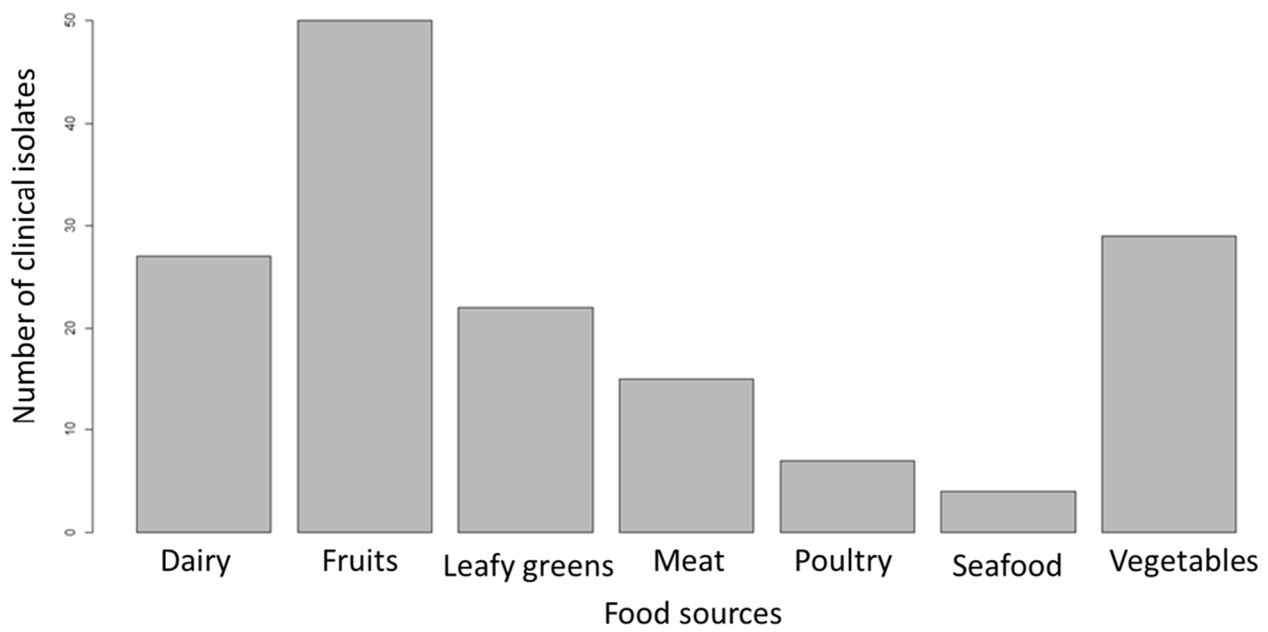

Figure 1. Predicted sources of clinical *L. monocytogenes* isolates.

2.3. Important Predictor Genes

Twenty of the most important genes were analyzed in isolates from different sources of food using logit boost, and their functional classes were determined based on an extensive literature survey. These genes allow us to identify microbial genetic patterns associated with each food source. According to Table 2, genes associated with survival, adaptation, and stress response were mainly found to be important in isolates from fresh produce, meat, and poultry. Additionally, two-component transcriptional regulators and virulence genes were found in isolates from fresh produce. However, some significant predictors/genes remained undefined in isolates from all food sources.

Table 2. Twenty putative genes sorted by maximum importance across the food sources.

Loci	Gene	Protein Name	Dairy	Fruits	Leafy Greens	Meat	Poultry	Seafood	Vegetables
lmo2702	*recR*	Recombination protein RecR	0.6653	0.5945	0.6925	0.8315	0.7212	0.6219	0.6653
lmo2401	*lmo2401*	Hypothetical protein	0.7017	0.663	0.6997	0.8231	0.7664	0.663	0.7017
lmo2615	*rpsE*	30S ribosomal protein S5	0.6873	0.5786	0.708	0.8199	0.7465	0.6081	0.6873
lmo2577	*lmo2577*	Hypothetical protein	0.7066	0.6611	0.6809	0.808	0.7851	0.6611	0.7066
lmo1501	*lmo1501*	Hypothetical protein	0.6925	0.6014	0.6839	0.8022	0.716	0.6374	0.6925
lmo1933	*folE*	GTP cyclohydrolase 1	0.577	0.6111	0.599	0.8012	0.7435	0.627	0.6111
lmo2215	*lmo2215*	Similar to ABC transporter (ATP-binding protein)	0.692	0.6473	0.6633	0.7988	0.72	0.6473	0.692
lmo0821	*lmo0821*	Hypothetical protein	0.6641	0.6641	0.7076	0.7979	0.7461	0.6641	0.657
lmo1715	*lmo1715*	Methyltransferase	0.674	0.6314	0.6612	0.7963	0.7371	0.6314	0.674
lmo2515	*degU*	NarL family, response regulator DegU	0.6923	0.6482	0.6759	0.7952	0.7781	0.6482	0.6923
lmo0625	*lmo0625*	Putative lipase/acylhydrolase	0.6548	0.6242	0.6813	0.7945	0.743	0.6242	0.6548
lmo0544	*srlA*	PTS sorbitol transporter subunit IIC	0.7125	0.6483	0.7073	0.7928	0.7713	0.6483	0.7125
lmo2728	*mlrA*	Transcriptional regulator, MerR family protein	0.62	0.6322	0.6294	0.7909	0.6994	0.6041	0.6322

Table 2. *Cont.*

Loci	Gene	Protein Name	Dairy	Fruits	Leafy Greens	Meat	Poultry	Seafood	Vegetables
lmo2348	*lmo2348*	Amino acid ABC transporter permease	0.6776	0.6673	0.681	0.7901	0.7512	0.6673	0.6776
lmo2422	*cesR*	Two-component response regulator	0.6988	0.6498	0.6574	0.7883	0.7307	0.6498	0.6988
lmo0623	*lmo0623*	Hypothetical protein	0.6382	0.6382	0.6382	0.7877	0.7026	0.6382	0.6307
lmo0635	*lmo0635*	Hypothetical protein	0.6715	0.6715	0.7008	0.7872	0.744	0.6715	0.656
lmo2658	*lmo2658*	Hypothetical protein	0.5621	0.5409	0.5969	0.7859	0.6298	0.5644	0.5621
lmo0611	*azoR1*	Azoreductase	0.626	0.6511	0.7853	0.7737	0.626	0.626	0.6511
lmo1425	*lmo1425*	Hypothetical protein	0.7079	0.651	0.6755	0.7852	0.7607	0.651	0.7079

Note: The numbers represent importance based on the accuracies of source prediction by each feature (genes). These values are the area under the receiver operating characteristic curve (AUC-ROC) determined from source-specific sensitivities and specificities (Supplementary Tables S1 and S2).

3. Discussion

3.1. Source Attribution Model

A major prerequisite for improving public health is preventing the emergence and spread of foodborne diseases. Source attribution models help link sporadic human cases of a specific foodborne illness to its food source. With the increasing usage of genome sequencing technologies, it is possible to identify genetic patterns indicative of the food source of pathogens. Recently, machine learning models have been used to identify molecular markers from foodborne pathogens linked with different hosts/phenotypes, which could be used to trace the source of human infections [26,36,37,39]. In the current study, we investigated the potential of machine learning to predict the food source origins of bacterial strains isolated from human cases of listeriosis using machine learning analyses of cgMLST data. Our machine learning model was able to recognize patterns in the complex dataset and use this information to predict the source of human listeriosis isolates. These patterns were based on variations in the genetic composition of *L. monocytogenes* isolated from different food sources. Furthermore, we identified allele variations that can be considered as being important predictors for this traceback process.

Due to the rapid adoption of genome sequencing technologies such as whole-genome sequencing (WGS) in food microbiology and public safety, new source attribution modeling approaches incorporating molecular information have been emerging. These methods generate comprehensive genomic data, providing critical insight into the transmission patterns of several major foodborne diseases, including listeriosis [42,43]. Here, we developed a machine learning-based source attribution model using the core genomes of 1212 *L. monocytogenes* isolates from different food sources. In our study, we have employed a high cutoff for the cgMLST allele calls. As a result, missing values in the cgMLST profiles can range from very low to very high, as seen in a prior study conducted by Kshirsagar and colleagues (2012). Another potential reason for missing data could be that some of the isolates may not possess the loci altogether. However, for successful modeling using machine learning techniques, complete data is essential, since missing values impact the overall effectiveness of the model(s). This issue can be overcome by imputing missing values [44]. In this study, missing allelic values in the food and clinical isolates were imputed. The total number of allele calls imputed in the isolates ranged from <1–78%, based on data completeness, which is consistent with that seen in a previous study [45]. As a result, our model performance was considerably improved, as seen in the model statistics. As shown in Table 1, logit boost was the best performing model (accuracy = 0.732, 95% confidence interval (CI) 0.665–0.760; Kappa = 0.654). A recent study [46] used a similar method to trace the source of salmonellosis, using random forest to determine the possible source of zoonotic outbreaks [46].

After testing a number of ML approaches, logit boost was used in source attribution in this study. Our model predicted that most of the listeriosis cases may have originated from

produce (fruits 32.5%; vegetables 18.8%; leafy greens 14.3%), 9.7% from meat, 4.6% from poultry, and 2.6% from seafood. Several studies have reported listeriosis outbreaks linked to the consumption of meats, dairy products, fresh produce, and seafood contaminated with *L. monocytogenes* [6,8,15,16,46–53]. Contamination of food sources may occur at any point in the production chain due to many factors [16,54]. The primary source of contamination or cross-contamination has been identified as originating from the farm environment, machinery, and staff [55–57]. This, however, is contingent on food handlers' level of hygienic practice. To avoid cross-contamination or recontamination during production and along the supply chain, food handlers must maintain personal hygiene and properly sanitize touch surfaces and production lines [58]. Finally, optimal cooking temperatures for specific food products should be considered during preparations [57], and temperatures in storage refrigerators should be properly monitored to prevent pathogens from growing, especially as *L. monocytogenes* stress response mechanisms allow it to survive non-thermal hurdle interventions [59–61].

3.2. Important Top Twenty Predictor Genes

Identifying the origin, also known as attribution, of microbial isolates is important within the realm of infectious diseases, specifically those caused due to direct or indirect contact with food or food sources. Prior efforts in this direction have focused on comparing the genotype, and its associated markers, of the isolate of interest with those seen in source populations [62–67]. Thus, it stands to reason that the increase in usage of genome sequencing methods in various aspects of food and outbreak surveillance should provide researchers with a wealth of features to analyze for source attribution purposes. However, the addition of such a large number of features can overwhelm current models due to the sheer scale of data and the amount of computation time added [62–67].

Prior studies have shown how such issues can be addressed by analyzing these complex data sets with ensemble machine learning classification [26,68]. In addition to accurate predictions, machine learning models can identify features that have the best prediction potential. Using our logit boost model, we identified 20 of the 736 *L. monocytogenes* genes that were the most important predictors for the attribution of listeriosis to different food sources. Our results (Table 2) showed that most of these genes were associated with *L. monocytogenes'* survival and stress response.

L. monocytogenes can adapt to, and survive, a wide range of stress conditions, including extremes of pH, temperature, and salt concentrations, which makes it problematic for food producers who rely on pathogen response to these stresses for food preservation. Stress tolerance in *L. monocytogenes* can be partially explained by the presence of the general stress response genes; transcription of these genes during host contamination provides homeostatic and protective functions to cope with the stress [11,69]. The *recR* gene, which encodes recombination protein and is involved in DNA repair, transcriptional genes *degU*, *cesR*, and *mlrA*, which encode putative response regulators that control many virulence factors, transporters *lmo2215* and *srlA*, and many genes coding for hypothetical proteins (*lmo2401*, *lmo2577*, *lmo2348*, *lmo0623*, *lmo0635*, *lmo2658*, and *lmo1425*) were identified as being important in association with the food sources studied. The putative DegU response regulator is a pleiotropic regulator involved in microbial motility at low temperatures [70]. This indicates the relevance of DegU in the current model, as most of the food sources studied are refrigerated or frozen to extend their shelf life—DegU may enable the survival of *L. monocytogenes* at low temperatures, contributing to its persistence in these foods, subsequently leading to listeriosis in humans who consume the contaminated food.

Furthermore, the presence of the response regulator CesR and the histidine protein kinase CesK, which is encoded by the gene downstream from *cesR*, indicates *L. monocytogenes'* ability to tolerate ethanol and antibiotics of the beta-lactam family (which act on the microbial cell wall) [71]. These genes may also enhance the persistence of *L. monocytogenes* in different food sources. Eight out of the twenty most important genes were hypothetical genes, which is in line with the findings of prior studies [36,39]. Thus, future studies in-

volving the characterization of each gene to understand its importance in *L. monocytogenes* adaptation and stress response along the food supply chain are warranted.

In the current study, we explored the use of machine learning in source attribution based on *L. monocytogenes* WGS data. Without a doubt, pathogens with food safety implications are not fully understood biologically, such as the relationship between specific infections and their sources. Our study shows that incorporating machine learning, surveillance, and monitoring infrastructures such as the National Antimicrobial Resistance Monitoring System and GenomeTrakr (which have been generating and uploading copious amounts of foodborne pathogen genomes) will allow researchers to draw a meaningful conclusion from genome-informed datasets. Machine learning is presumably positioned to address many of the current challenges in the food safety industry. By using machine learning, it may be possible to uncover patterns in WGS data that are not easily gleaned from traditional methods. Thus, it may be possible to solve difficult problems in food source attribution using genomic data.

In conclusion, supervised machine learning was effective in attributing food sources to listeriosis clinical cases based on WGS data. Inferring genetic information from pathogen genotypes often proves crucial for biological inference. Source attribution of *L. monocytogenes* infections allows food industry professionals, data managers, epidemiologists, microbiologists, and bioinformaticians to tailor their practices to prevent the spread of foodborne pathogens. It also enables healthcare professionals to more efficiently use resources to contain the survival and proliferation of pathogens at the source. As genomic data becomes more widely available, WGS serves as a cost-effective method for public health surveillance. With the availability of hundreds of thousands of genomes of foodborne pathogens and evolutionary relationships rapidly being determined, sequencing information can be used for prediction purposes when combined with useful isolate metadata, particularly in the food safety domain. One limitation of this study was that, while an ideal validation scenario would involve validating the model on a new data set (such as an unused subsample of data during model training), all of our data were used for model development due to the limited number of samples. However, in the future, the model can be validated on new data as it becomes available.

4. Materials and Methods

4.1. Data Description

L. Monocytogenes isolates included in this study (n = 1366) were sampled from the National Centers for Biotechnology Information's (NCBI) Pathogen Detection database and included 154 isolates from human listeriosis patients and 1212 isolates from food sources, including dairy (197), fruits (302), leafy greens (115), meat (119), poultry (119), seafood (145), and vegetables (215) (Supplementary Table S3). The included *L. monocytogenes* isolates were extracted from food and clinical sources as part of integrated surveillance and were previously sequenced (using different platforms such as Illumina HiSeq, NextSeq, or MiSeq). A simple random sampling of 10 to 60% of all available isolates from each source was performed, based on the availability of relevant metadata (such as location, isolation source, source type, and Interagency Food Safety Analytics Collaboration (IFSAC) category), which served as isolate inclusion criteria. The clinical isolates selected were also sampled from publicly available sequences, and as such were not epidemiologically associated with specific outbreaks.

4.2. Bioinformatics Analysis

Input for the source attribution model was generated from all sequences within the data set by running cgMLST. The Enterobase scheme was used to obtain cgMLST [72,73] in BioNumerics v.7.6 (Applied Maths, Sint Martens Latem, Belgium). *L. monocytogenes* has 1748 core genes, with each loci having several allele variations [72,73]. The cgMLST allele calls were accepted when the strains had a core genome coverage of more than 95% (1661) of the 1748 core genome alleles, and detection of mixed sequence alleles of less than

50 alleles. In some cases, BioNumerics may fail to call an allele as a result of stop codons, indels, and other factors in the genome sequence, resulting in missing values in the cgMLST profile. In such cases, we used the missForest package in R (version 4.1.2) to impute the missing values. In the missForest package, missing values are imputed using random forest trained on the observed data to predict the missing values [55].

4.3. Source Attribution Modeling

Machine learning algorithms were used to predict the food source of a given strain isolated from human listeriosis cases based on allelic variations found in the core genes of *L. monocytogenes* isolated from food sources (dairy, fruits, green leafy vegetables, meat, poultry, and seafood). In this study, we used supervised machine learning classification models. Here, our models learned patterns in the allelic variations of the *L. monocytogenes* isolates from food sources. Modeling was carried out in R (v. 4.1.2, R Core Team, 2021; Vienna, Austria) using the caret package [74,75].

4.3.1. Feature Reduction

The core genome of *L. monocytogenes* consists of 1748 loci [72,73]. Feature reduction was performed using the nearZeroVar (near zero variance) function in the caret package in R to remove some features (genes). NearZeroVar identifies features that have a single unique value or have very few unique values relative to the number of samples, or when the frequency ratio (frequency of most frequent value divided by the frequency of second most frequent value) is large [74]. This is because retaining these redundant features that provide no useful details to distinguish between the food sources may only increase computation time and model complexity.

4.3.2. Machine Learning

Our feature-reduced cgMLST data was randomly split into a training set (70%) and a testing set (30%). Four machine learning algorithms–random forest, logit boost, stochastic gradient boosting, and support vector machine radial kernel–have been successfully applied in studies analyzing WGS data [36,37,39,76] and were therefore used in training our data. We used 10-fold cross-validation, which randomly partitions the training data set into 10 equal folds—nine folds used for model training and one fold to estimate model performance—to obtain the model with the best performance. This procedure was repeated until 10 models had been trained, each using unique training and testing folds. The default hyperparameter grid (in the R package *caret*) was employed to search for optimal tuning parameters for all four algorithms. The final tuning parameters utilized for the models, based on the best-fit Kappa values, were: LB (31 nIter), RF (38 mtry), GBM (150 n.trees, 3 interaction depth, 0.1 shrinkage, and 10 n.minobsinnode), and SVMR (0.002580397 sigma and 1 C).

The developed models were evaluated against the testing set and the performance of the models was assessed based on the Kappa value, model accuracy, and other confusion matrix statistics. The accuracy was calculated from the models' ability to correctly classify the testing data set. The Kappa value is a statistic that compares the model accuracy (observed accuracy) with the expected accuracy [77]. It shows the agreement between predicted and actual classes and is especially important in highly unbalanced data where the accuracy can be misleading. We performed 10 iterations of model training and testing and selected the algorithm that achieved the highest average accuracy as the best algorithm for further analysis.

A final model was developed by training the best-performing algorithm on the complete feature-reduced cgMLST data. This was run to allow the algorithm to learn as much as possible from the variability in the complete data set. This approach has been successfully implemented by [26] and has been identified as the best approach for a predictive model. The best performing model was then used to predict the probable food sources of each of the 154 clinical *L. monocytogenes* isolates.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pathogens11060691/s1, Table S1: Statistics of the confusion matrices for LB, RF, GBM, and SVMR (train and test models); Table S2: Statistics of the confusion matrices for the final (best performing) model (LB); Table S3: *Listeria monocytogenes* isolates (indicated by their BioSample numbers) from food sources and clinical samples used to generate cgMLST profiles.

Author Contributions: Conceptualization, A.K.P. and C.K.T.; methodology, C.K.T.; validation, C.K.T., A.K.P., E.O.B. and S.K.; formal analysis, C.K.T., E.O.B. and S.K.; investigation, C.K.T., E.O.B. and S.K.; resources, A.K.P.; data curation, C.K.T. and E.O.B.; writing—original draft preparation, C.K.T.; writing—review and editing, C.K.T., A.K.P., E.O.B. and S.K.; supervision, A.K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. CDC. *CDC Estimates of Foodborne Illness in the United States*; CDC: Atlanta, GA, USA, 2018; Volume 68.
2. Scallan, E.; Hoekstra, R.M.; Angulo, F.J.; Tauxe, R.V.; Widdowson, M.A.; Roy, S.L.; Jones, J.L.; Griffin, P.M. Foodborne illness acquired in the United States-Major pathogens. *Emerg. Infect. Dis.* **2011**, *17*, 7–15. [CrossRef] [PubMed]
3. Gourama, H. Foodborne Pathogens. In *Food Engineering Series*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 25–49.
4. Dewey-Mattia, D.; Manikonda, K.; Hall, A.J.; Wise, M.E.; Crowe, S.J. Surveillance for foodborne disease outbreaks—United States, 2009–2015. *MMWR Surveill. Summ.* **2018**, *67*, 1–11. [CrossRef] [PubMed]
5. Painter, J.A.; Hoekstra, R.M.; Ayers, T.; Tauxe, R.V.; Braden, C.R.; Angulo, F.J.; Griffin, P.M. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008. *Emerg. Infect. Dis.* **2013**, *19*, 407–415. [CrossRef] [PubMed]
6. Filipello, V.; Mughini-Gras, L.; Gallina, S.; Vitale, N.; Mannelli, A.; Pontello, M.; Decastelli, L.; Allard, M.W.; Brown, E.W.; Lomonaco, S. Attribution of Listeria monocytogenes human infections to food and animal sources in Northern Italy. *Food Microbiol.* **2020**, *89*, 103433. [CrossRef]
7. Riley, L.W. Extraintestinal foodborne pathogens. *Annu. Rev. Food Sci. Technol.* **2020**, *11*, 275–294. [CrossRef]
8. CDC. Listeria (Listeriosis) | Listeria | CDC. Available online: https://www.cdc.gov/listeria/index.html (accessed on 4 May 2021).
9. Chlebicz, A.; Śliżewska, K. Campylobacteriosis, salmonellosis, yersiniosis, and listeriosis as zoonotic foodborne diseases: A review. *Int. J. Environ. Res. Public Health* **2018**, *15*, 863. [CrossRef]
10. Lomonaco, S.; Nucera, D.; Filipello, V. The evolution and epidemiology of Listeria monocytogenes in Europe and the United States. *Infect. Genet. Evol.* **2015**, *35*, 172–183. [CrossRef]
11. Vivant, A.L.; Garmyn, D.; Piveteau, P. Listeria monocytogenes, a down-to-earth pathogen. *Front. Cell. Infect. Microbiol.* **2013**, *3*, 87. [CrossRef]
12. FDA. Get the Facts about Listeria | FDA. Available online: https://www.fda.gov/animal-veterinary/animal-health-literacy/get-facts-about-listeria# (accessed on 7 May 2022).
13. Lopez-Valladares, G.; Danielsson-Tham, M.L.; Tham, W. Implicated food products for listeriosis and changes in serovars of Listeria monocytogenes affecting humans in recent decades. *Foodborne Pathog. Dis.* **2018**, *15*, 387–397. [CrossRef]
14. Heredia, N.; García, S. Animals as sources of food-borne pathogens: A review. *Anim. Nutr.* **2018**, *4*, 250–255. [CrossRef]
15. Maćkiw, E.; Korsak, D.; Kowalska, J.; Felix, B.; Stasiak, M.; Kucharek, K.; Postupolski, J. Incidence and genetic variability of Listeria monocytogenes isolated from vegetables in Poland. *Int. J. Food Microbiol.* **2021**, *339*, 109023. [CrossRef]
16. Marik, C.M.; Zuchel, J.; Schaffner, D.W.; Strawn, L.K. Growth and survival of Listeria monocytogenes on intact fruit and vegetable surfaces during postharvest handling: A systematic literature review. *J. Food Prot.* **2020**, *83*, 108–128. [CrossRef]
17. Matthews, K.; Kniel, K.; Montville, T. *Food Microbiology: An Introduction*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
18. Batz, M.B.; Doyle, M.P.; Morris, J.G.; Painter, J.; Singh, R.; Tauxe, R.V.; Taylor, M.R.; Wong, D.M.; Food Attribution Working Group. Attributing illness to food. *Emerg. Infect. Dis.* **2005**, *11*, 993–999. [CrossRef]
19. Hoffmann, S.; Maculloch, B.; Batz, M. Economic burden of major foodborne illnesses acquired in the United States. In *Economic Cost of Foodborne Illnesses in the United States*; EIB-140; U.S. Department of Agriculture, Economic Research Service: Washington, DC, USA, 2015; pp. 1–74. ISBN 9781634836661. Available online: https://www.ers.usda.gov/publications/pub-details/?pubid=43987 (accessed on 7 May 2022).

20. Franz, E.; Gras, L.M.; Dallman, T. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Curr. Opin. Food Sci.* **2016**, *8*, 74–79. [CrossRef]
21. Pires, S.M.; Evers, E.G.; Van Pelt, W.; Ayers, T.; Scallan, E.; Angulo, F.J.; Havelaar, A.; Hald, T.; Schroeter, A.; Brisabois, A.; et al. Attributing the human disease burden of foodborne infections to specific sources. *Foodborne Pathog. Dis.* **2009**, *6*, 417–424. [CrossRef]
22. Todd, E.C.D.; Notermans, S. Surveillance of listeriosis and its causative pathogen, Listeria monocytogenes. *Food Control* **2011**, *22*, 1484–1490. [CrossRef]
23. de Knegt, L.V.; Pires, S.M.; Löfström, C.; Sørensen, G.; Pedersen, K.; Torpdahl, M.; Nielsen, E.M.; Hald, T. Application of molecular typing results in source attribution models: The case of multiple locus variable number tandem repeat analysis (MLVA) of Salmonella isolates obtained from integrated surveillance in Denmark. *Risk Anal.* **2016**, *36*, 571–588. [CrossRef]
24. Mughini-Gras, L.; Kooh, P.; Augustin, J.C.; David, J.; Fravalo, P.; Guillier, L.; Jourdan-Da-Silva, N.; Thébault, A.; Sanaa, M.; Watier, L.; et al. Source attribution of foodborne diseases: Potentialities, hurdles, and future expectations. *Front. Microbiol.* **2018**, *9*, 1983. [CrossRef]
25. Dingle, K.E.; Colles, F.M.; Wareing, D.R.A.; Ure, R.; Fox, A.J.; Bolton, F.E.; Bootsma, H.J.; Willems, R.J.L.; Urwin, R.; Maiden, M.C.J. Multilocus sequence typing system for Campylobacter jejuni. *J. Clin. Microbiol.* **2001**, *39*, 14–23. [CrossRef]
26. Munck, N.; Njage, P.M.K.; Leekitcharoenphon, P.; Litrup, E.; Hald, T. Application of whole-genome sequences and machine learning in source attribution of Salmonella Typhimurium. *Risk Anal.* **2020**, *40*, 1693–1705. [CrossRef]
27. Sheppard, S.K.; Dallas, J.F.; MacRae, M.; McCarthy, N.D.; Sproston, E.L.; Gormley, F.J.; Strachan, N.J.C.; Ogden, I.D.; Maiden, M.C.J.; Forbes, K.J. Campylobacter genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. *Int. J. Food Microbiol.* **2009**, *134*, 96–103. [CrossRef] [PubMed]
28. Mullner, P.; Spencer, S.E.F.; Wilson, D.J.; Jones, G.; Noble, A.D.; Midwinter, A.C.; Collins-Emerson, J.M.; Carter, P.; Hathaway, S.; French, N.P. Assigning the source of human campylobacteriosis in New Zealand: A comparative genetic and epidemiological approach. *Infect. Genet. Evol.* **2009**, *9*, 1311–1319. [CrossRef] [PubMed]
29. Kurpas, M.; Osek, J.; Moura, A.; Leclercq, A.; Lecuit, M.; Wieczorek, K. Genomic Characterization of Listeria monocytogenes isolated from ready-to-eat meat and meat processing environments in Poland. *Front. Microbiol.* **2020**, *11*, 1412. [CrossRef] [PubMed]
30. Jagadeesan, B.; Baert, L.; Wiedmann, M.; Orsi, R.H. Comparative analysis of tools and approaches for source tracking Listeria monocytogenes in a food facility using whole-genome sequence data. *Front. Microbiol.* **2019**, *10*, 947. [CrossRef]
31. Foley, S.L.; Johnson, T.J.; Ricke, S.C.; Nayak, R.; Danzeisen, J. Salmonella Pathogenicity and host adaptation in chicken-associated serovars. *Microbiol. Mol. Biol. Rev.* **2013**, *77*, 582–607. [CrossRef]
32. Monack, D.M. Salmonella persistence and transmission strategies. *Curr. Opin. Microbiol.* **2012**, *15*, 100–107. [CrossRef]
33. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [CrossRef]
34. Baker, R.E.; Peña, J.M.; Jayamohan, J.; Jérusalem, A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.* **2018**, *14*, 20170660. [CrossRef]
35. Alkema, W.; Boekhorst, J.; Wels, M.; Van Hijum, S.A.F.T. Microbial bioinformatics for food safety and production. *Brief. Bioinform.* **2016**, *17*, 283–292. [CrossRef]
36. Tanui, C.K.; Karanth, S.; Njage, P.M.K.; Meng, J.; Pradhan, A.K. Machine learning-based predictive modeling to identify genotypic traits associated with Salmonella enterica disease endpoints in isolates from ground chicken. *LWT* **2022**, *154*, 112701. [CrossRef]
37. Njage, P.M.K.; Henri, C.; Leekitcharoenphon, P.; Mistou, M.Y.; Hendriksen, R.S.; Hald, T. Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. *Risk Anal.* **2019**, *39*, 1397–1413. [CrossRef]
38. Lupolova, N.; Dallman, T.J.; Holden, N.J.; Gally, D.L. Erratum: Patchy promiscuity: Machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. *Microb. Genom.* **2018**, *4*, e000193. [CrossRef]
39. Karanth, S.; Tanui, C.K.; Meng, J.; Pradhan, A.K. Exploring the predictive capability of advanced machine learning in identifying severe disease phenotype in Salmonella enterica. *Food Res. Int.* **2022**, *151*, 110817. [CrossRef]
40. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159. [CrossRef]
41. Fleiss, J.L.; Levin, B.; Paik, M.C. *Statistical Methods for Rates and Proportions*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2003; ISBN 0471526290.
42. Cabal, A.; Pietzka, A.; Huhulescu, S.; Allerberger, F.; Ruppitsch, W.; Schmid, D. Isolate-based surveillance of Listeria monocytogenes by whole genome sequencing in Austria. *Front. Microbiol.* **2019**, *10*, 2282. [CrossRef]
43. Chen, J.; Karanth, S.; Pradhan, A.K. Quantitative microbial risk assessment for Salmonella: Inclusion of whole genome sequencing and genomic epidemiological studies, and advances in the bioinformatics pipeline. *J. Agric. Food Res.* **2020**, *2*, 100045. [CrossRef]
44. Stekhoven, D.J.; Bühlmann, P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
45. Kshirsagar, M.; Carbonell, J.; Klein-Seetharaman, J. Techniques to cope with missing data in host-pathogen protein interaction prediction. *Bioinformatics* **2012**, *28*, 466–472. [CrossRef]

46. Zhang, S.; Li, S.; Gu, W.; Den Bakker, H.; Boxrud, D.; Taylor, A.; Roe, C.; Driebe, E.; Engelthaler, D.M.; Allard, M.; et al. Zoonotic source attribution of Salmonella enterica serotype Typhimurium using genomic surveillance data, United States. *Emerg. Infect. Dis.* **2019**, *25*, 82–91. [CrossRef]
47. Mritunjay, S.K.; Kumar, V. Fresh farm produce as a source of pathogens: A review. *Res. J. Environ. Toxicol.* **2015**, *9*, 59–70. [CrossRef]
48. Aureli, P.; Fiorucci, G.C.; Caroli, D.; Marchiaro, G.; Novara, O.; Leone, L.; Salmaso, S. An outbreak of febrile gastroenteritis associated with corn contaminated by Listeria monocytogenes. *N. Engl. J. Med.* **2000**, *342*, 1236–1241. [CrossRef]
49. Angelo, K.M.; Conrad, A.R.; Saupe, A.; Dragoo, H.; West, N.; Sorenson, A.; Barnes, A.; Doyle, M.; Beal, J.; Jackson, K.A.; et al. Multistate outbreak of Listeria monocytogenes infections linked to whole apples used in commercially produced, prepackaged caramel apples: United States, 2014-2015. *Epidemiol. Infect.* **2017**, *145*, 848–856. [CrossRef]
50. Zilelidou, E.A.; Tsourou, V.; Poimenidou, S.; Loukou, A.; Skandamis, P.N. Modeling transfer of Escherichia coli O157: H7 and Listeria monocytogenes during preparation of fresh-cut salads: Impact of cutting and shredding practices. *Food Microbiol.* **2015**, *45*, 254–265. [CrossRef]
51. Norton, D.M.; Braden, C.R. Foodborne listeriosis. In *Listeria, Listeriosis, and Food Safety, Third Edition*; CRC Press Taylor & Francis Group: Boca Raton, FL, USA, 2007; pp. 305–356. ISBN 9781420015188.
52. Mashak, Z.; Banisharif, F.; Banisharif, G.; Reza Pourian, M.; Eskandari, S.; Seif, A.; Safarpoor Dehkordi, F.; Alavi, I. Prevalence of Listeria species and serotyping of Listeria monocytogenes bacteria isolated from seafood samples. *Egypt. J. Vet. Sci.* **2021**, *52*, 1–9. [CrossRef]
53. CDC. *National Outbreak Reporting System*; CDC: Atlanta, GA, USA, 2019.
54. Gil, M.I.; Selma, M.V.; Suslow, T.; Jacxsens, L.; Uyttendaele, M.; Allende, A. Pre- and postharvest preventive measures and intervention strategies to control microbial food safety hazards of fresh leafy vegetables. *Crit. Rev. Food Sci. Nutr.* **2015**, *55*, 453–468. [CrossRef] [PubMed]
55. Osaili, T.M.; Alaboudi, A.R.; Nesiar, E.A. Prevalence of Listeria spp. and antibiotic susceptibility of Listeria monocytogenes isolated from raw chicken and ready-to-eat chicken products in Jordan. *Food Control* **2011**, *22*, 586–590. [CrossRef]
56. Schäfer, D.F.; Steffens, J.; Barbosa, J.; Zeni, J.; Paroul, N.; Valduga, E.; Junges, A.; Backes, G.T.; Cansian, R.L. Monitoring of contamination sources of Listeria monocytogenes in a poultry slaughterhouse. *LWT* **2017**, *86*, 393–398. [CrossRef]
57. Carrasco, E.; Morales-Rueda, A.; García-Gimeno, R.M. Cross-contamination and recontamination by Salmonella in foods: A review. *Food Res. Int.* **2012**, *45*, 545–556. [CrossRef]
58. Bogere, P.; Baluka, A.S. Microbiological quality of meat at the abattoir and butchery levels in Kampala city, Uganda. *Internet J. Food Saf.* **2014**, *16*, 29–35.
59. Lambertz, S.T.; Nilsson, C.; Brådenmark, A.; Sylvén, S.; Johansson, A.; Jansson, L.M.; Lindblad, M. Prevalence and level of Listeria monocytogenes in ready-to-eat foods in Sweden 2010. *Int. J. Food Microbiol.* **2012**, *160*, 24–31. [CrossRef]
60. Matle, I.; Mbatha, K.R.; Lentsoane, O.; Magwedere, K.; Morey, L.; Madoroba, E. Occurrence, serotypes, and characteristics of Listeria monocytogenes in meat and meat products in South Africa between 2014 and 2016. *J. Food Saf.* **2019**, *39*, e12629. [CrossRef]
61. Vitas, A.I.; Garcia-Jalon, V.A.E.I. Occurrence of Listeria monocytogenes in fresh and processed foods in Navarra (Spain). *Int. J. Food Microbiol.* **2004**, *90*, 349–356. [CrossRef]
62. McCarthy, N.D.; Colles, F.M.; Dingle, K.E.; Bagnall, M.C.; Manning, G.; Maiden, M.C.J.; Falush, D. Host-associated genetic import in Campylobacter jejuni. *Emerg. Infect. Dis.* **2007**, *13*, 267–272. [CrossRef]
63. Sheppard, S.K.; Dallas, J.F.; Strachan, N.J.C.; MacRae, M.; McCarthy, N.D.; Wilson, D.J.; Gormley, F.J.; Falush, D.; Ogden, L.D.; Maiden, M.C.J.; et al. Campylobacter genotyping to determine the source of human infection. *Clin. Infect. Dis.* **2009**, *48*, 1072–1078. [CrossRef]
64. Strachan, N.J.C.; Gormley, F.J.; Rotariu, O.; Ogden, I.D.; Miller, G.; Dunn, G.M.; Sheppard, S.K.; Dallas, J.F.; Reid, T.M.S.; Howie, H.; et al. Attribution of campylobacter infections in Northeast Scotland to specific sources by use of multilocus sequence typing. *J. Infect. Dis.* **2009**, *199*, 1205–1208. [CrossRef]
65. Rosner, B.M.; Schielke, A.; Didelot, X.; Kops, F.; Breidenbach, J.; Willrich, N.; Gölz, G.; Alter, T.; Stingl, K.; Josenhans, C.; et al. A combined case-control and molecular source attribution study of human Campylobacter infections in Germany, 2011–2014. *Sci. Rep.* **2017**, *7*, 5139. [CrossRef]
66. Miller, P.; Marshall, J.; French, N.; Jewell, C. sourceR: Classification and source attribution of infectious agents among heterogeneous populations. *PLoS Comput. Biol.* **2017**, *13*, e1005564. [CrossRef]
67. Maiden, M.C.J.; Van Rensburg, M.J.J.; Bray, J.E.; Earle, S.G.; Ford, S.A.; Jolley, K.A.; McCarthy, N.D. MLST revisited: The gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **2013**, *11*, 728–736. [CrossRef]
68. Arning, N.; Sheppard, S.K.; Bayliss, S.; Clifton, D.A.; Wilson, D.J. Machine learning to predict the source of campylobacteriosis using whole genome data. *PLoS Genet.* **2021**, *17*, e1009436. [CrossRef]
69. Beales, N. Adaptation of microorganisms to cold temperatures, weak acid preservatives, low pH, and osmotic stress: A review. *Compr. Rev. Food Sci. Food Saf.* **2004**, *3*, 1–20. [CrossRef]
70. Knudsen, G.M.; Olsen, J.E.; Dons, L. Characterization of DegU, a response regulator in Listeria monocytogenes, involved in regulation of motility and contributes to virulence. *FEMS Microbiol. Lett.* **2004**, *240*, 171–179. [CrossRef]

71. Kallipolitis, B.H.; Ingmer, H.; Gahan, C.G.; Hill, C.; Søgaard-Andersen, L. CesRK, a Two-component signal transduction system in Listeria monocytogenes, responds to the presence of cell wall-acting antibiotics and affects β-Lactam resistance. *Antimicrob. Agents Chemother.* **2003**, *47*, 3421–3429. [CrossRef]
72. Moura, A.; Criscuolo, A.; Pouseele, H.; Maury, M.M.; Leclercq, A.; Tarr, C.; Björkman, J.T.; Dallman, T.; Reimer, A.; Enouf, V.; et al. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nat. Microbiol.* **2016**, *2*, 16185. [CrossRef]
73. Zhou, Z.; Alikhan, N.F.; Sergeant, M.J.; Luhmann, N.; Vaz, C.; Francisco, A.P.; Carriço, J.A.; Achtman, M. Grapetree: Visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.* **2018**, *28*, 1395–1404. [CrossRef] [PubMed]
74. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
75. Kuhn, M. The Caret Package. 2011. Available online: https://topepo.github.io/caret/ (accessed on 7 May 2022).
76. Wheeler, N.E.; Gardner, P.P.; Barquist, L. Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enterica. *PLoS Genet.* **2018**, *14*, e1007333. [CrossRef]
77. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; ISBN 9781461468493.