School of Public Health
BROWN UNIVERSITY

Statistical Machine Learning for Foodborne Disease Source Attribution

Rophence Ojiambo[1], Zexuan Yu[1], and Amos Okutse[1]
[1]Brown University, RI, Providence

Overview

We developed a statistical machine learning classification model for Listeria monocytogene pathogen food source attribution based on a comparative analysis of Naïve Bayes and Random Forest algorithms.

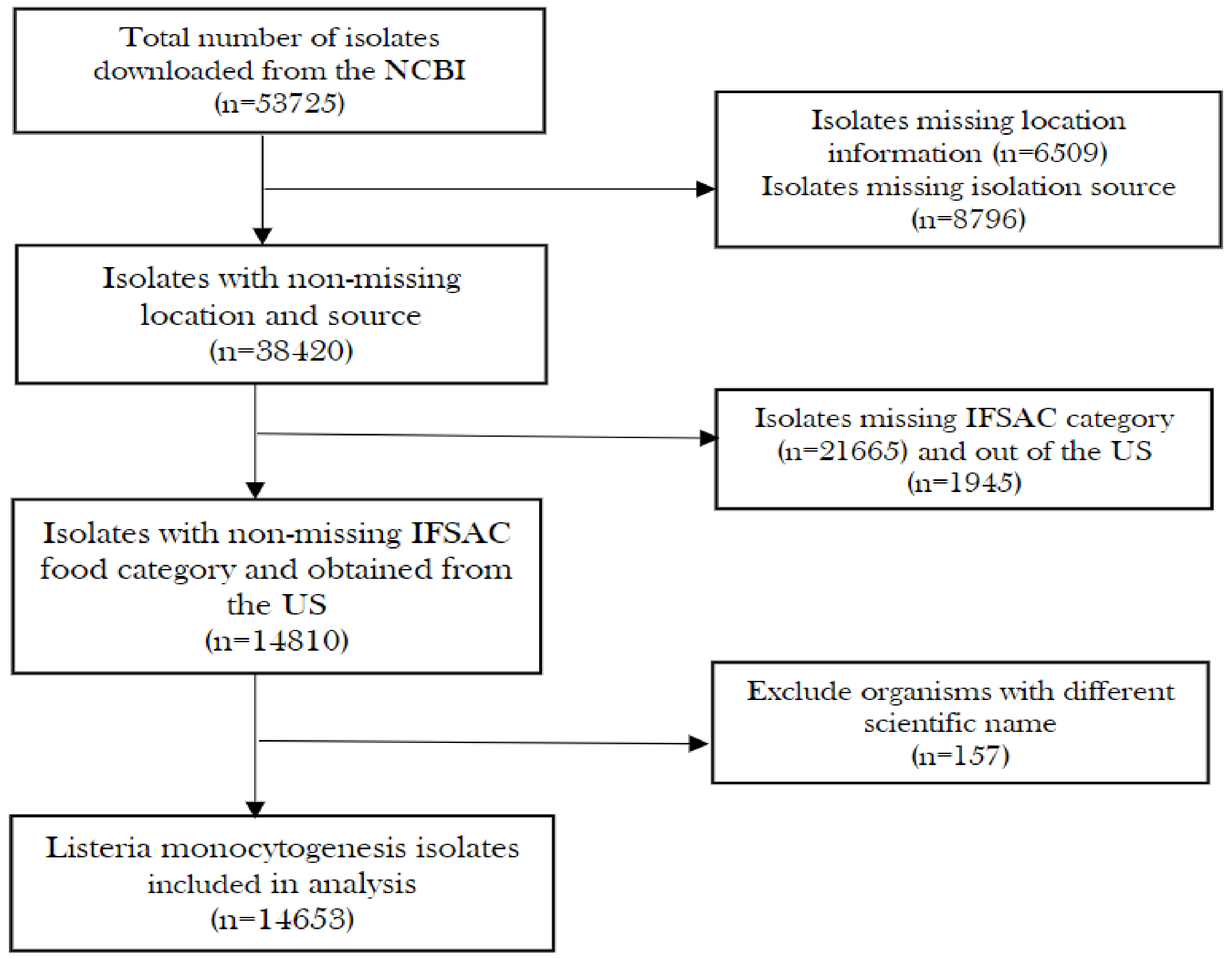
Background

- In the US, foodborne illnesses result in approximately 128,000 hospitalizations and 3,000 fatalities.
- The US Centers for Disease Control and Prevention (CDC) notes that approximately 1,600 cases of listeriosis are recorded annually with about 260 mortalities.
- Outbreak investigations have shown links between these pathogens and specific food sources.
- We compared the performance of Naïve Bayes and Random Forest algorithms to develop a model for foodborne-illness source attribution linked to L. monocytogene isolates.

Study Design and Methods

- Given strains of L. Monocytogenes isolates sampled from the National Centers for Biotechnology Information's (NCBI) Pathogen Detection database, Naïve Bayes and Random Forest modeling were used to predict food sources.
  - There were 66 isolates from human listeriosis patients, 5696 from environmental sources, 1967 from food sources and 618 isolates from other sources.
  - Food sources included dairy (624), fruits (258), leafy greens (70), meat (267), poultry (254), sea food (198) and vegetables (296).
  - Isolation source was the primary outcome while the features included Strain, Season, State, Single Nucleotide Polymorphism (SNP) clusters, Isolate, Min.same and Min.diff variables.
- Data was split into training and test sets (75% and 25% respectively).
  - A 10-fold cross validation and minority class up-sampling was employed to develop the Naïve Bayes and Random Forest (RF) machine learning algorithms in the training data.
  - Final tuning parameters were; α=1 Laplace smoothing for the Naïve Bayes; mtry=2 for the Random Forest model.
  - The developed models were evaluated against the test set and performance assessed based on model accuracy, Kappa values and other confusion matrix statistics.
  - The final model was developed by the training the best algorithm; the Random Forest model on the full training data.

Study Flowchart



Total number of isolates downloaded from the NCBI (n=53725)

Isolates missing location information (n=6509)
Isolates missing isolation source (n=8796)

Isolates with non-missing location and source (n=38420)

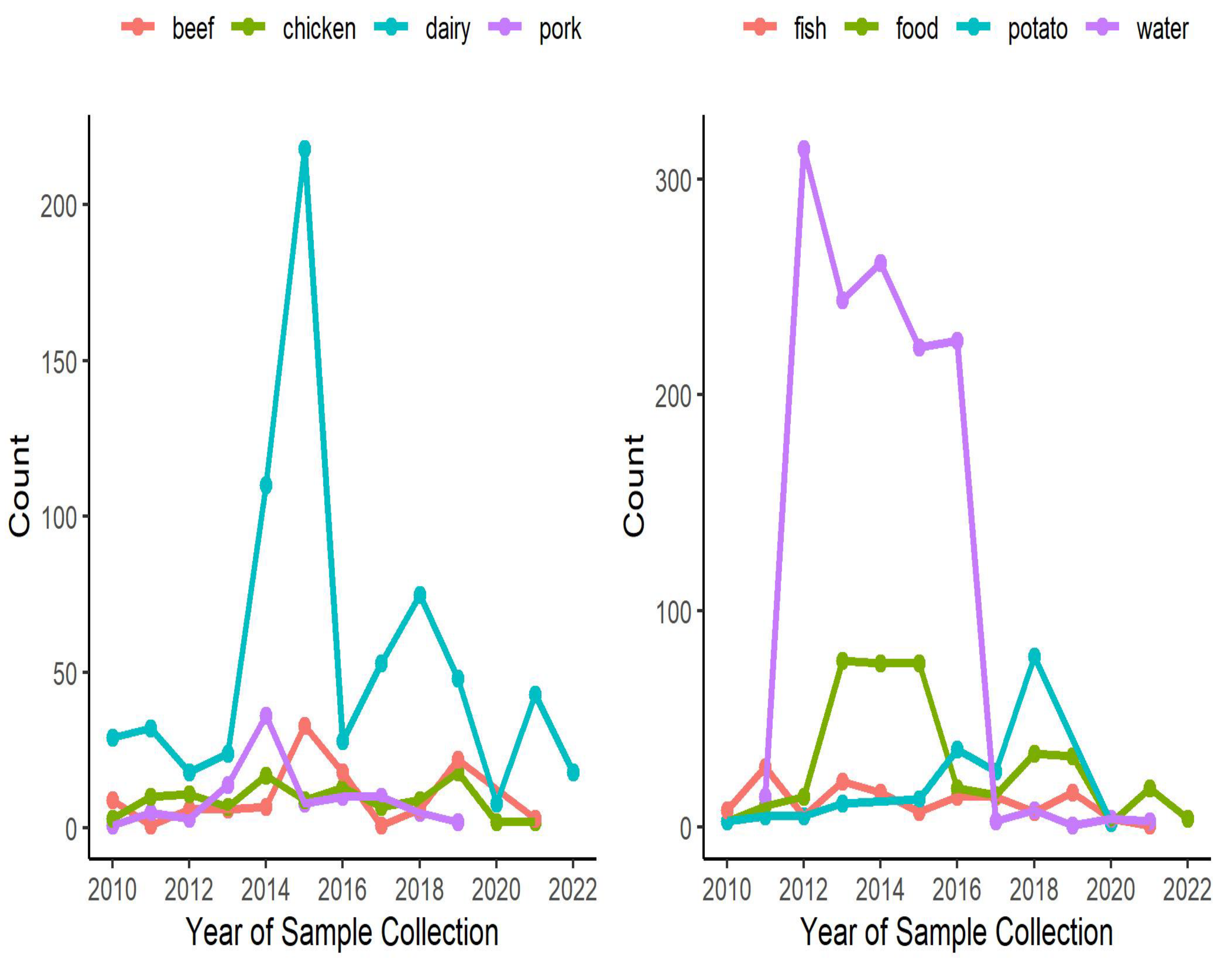
Isolates missing IFSAC category (n=21665) and out of the US (n=1945)

Isolates with non-missing IFSAC food category and obtained from the US (n=14810)

Exclude organisms with different scientific name (n=157)

Listeria monocytogenesis isolates included in analysis (n=14653)

Trends of L. monocytogenes counts by top isolation sources



Results

Table 1. Performance measures across 10 folds with resampling for Naive Bayes and random forest classification algorithms

Metric	Naïve Bayes		Random Forest	
	Estimate	Standard error (SE)	Estimate	Standard error (SE)
Accuracy	0.2577	0.0065	0.8727	0.0033
Jaccard's Index	0.2831	0.009	0.6519	0.0103
Kappa	0.1398	0.0046	0.7438	0.0055
AUC	0.7961	0.0057	0.9453	0.0028
Sensitivity	0.3633	0.0092	0.6771	0.0105
Specificity	0.9198	0.0007	0.9748	0.0006

Table 2. Naïve Bayes and RF Performance measures on test data

Metric	Naïve Bayes	Random Forest
Accuracy	0.2496	0.8625
Jaccard's Index	0.2515	0.5533
Kappa	0.132	0.7194
AUC	0.7597	0.9265
Sensitivity	0.3324	0.5813
Specificity	0.9191	0.972

Table 3. Performance measures of Random Forest on the full data

Metric	Random Forest	
	Estimate	Standard error (SE)
Accuracy	0.884	0.0036
Jaccard's Index	0.6874	0.0116
Kappa	0.7687	0.0043
AUC	0.958	0.0019
Sensitivity	0.7098	0.0115
Specificity	0.9776	0.0004

- The performance of Naïve Bayes and Random forest differed significantly from one another with an average accuracy of 0.258 and 0.873, respectively.
- Strain was the most important feature followed by Isolate, state, Min.diff, Min.same, season and SNP cluster variables, respectively.

Conclusion

- The Random forest model had a very good ability to discriminate between the different food sources, with an area under the curve (AUC) of 0.945 compared to the Naïve Bayes model that had moderately good discriminatory ability, AUC of 0.796.
- Statistical machine learning methods promise efficiency in food source attribution of L. monocytogenes isolates, which can substantially enhance investigation of outbreak cases and reduce the pool of food sources targeted by intervention policies.
- The study was restricted to a complete case analysis due to high percentage of missing data. Categorization of the isolation source variable may have introduced bias and loss of efficiency in the results.

References

National Library of Medicine (US) National Center for Biotechnology Information B (MD): The NCBI pathogen detection project [internet], https://www.ncbi.nlm.nih.gov/pathogens/ (2016).

Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. Chapman; Hall/CRC, 1995.

Breiman L. Random forests. Machine learning 2001; 45: 5–32