Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 3, Issue 8, 2025

IDENTIFICATION AND CLASSIFICATION OF FOODBORNE DISEASE OUTBREAKS

Ali Zain[*1], Asad Ali Zakir[2], Shehar Zaad[3], Saira Shairi[4], Qaiser Nadeem[5]

[*1]*University Of Engineering and Technology, Department of Computer Science, Lahore, Pakistan.*
[2,4,5]*Department of Computer Science, University of Central Punjab, Lahore, Pakistan.*
[3]*Department of Software Engineering, University of Central Punjab, Lahore, Pakistan.*

[*1]alizain15@live.com, [2]iamasadali66@gmail.com, [3]shehar.zaad@ucp.edu.pk, [4]saira.shairi@ucp.edu.pk, [5]qaisernadeem042@gmail.com

DOI:https://doi.org/10.5281/zenodo.16717546

Keywords
Classification, Foodborne, Outbreaks, Causative Agents, Naïve Bayes, Decision Tree, Random Forest.

Article History
Received on 11 May 2025
Accepted on 21 July 2025
Published on 02 August 2025

Corresponding Author: *
Ali Zain

Abstract
Foodborne disease is commonly caused by consuming contaminated food and beverages so the identification and classification of foodborne disease outbreaks is necessary to prevent and reduce the risk of illness and death. The purpose of this research is to identify the causative agents of disease as soon as possible to improve the food safety to prevent from illnesses and deaths. The useful patterns have been identified with analysis on dataset and also determine the large number of outbreaks occurs in year, food, location and species. The classification is done in Decision Tree, Naïve Bayes and Random Forest classifiers. The experiments on the dataset have proven the efficiency of purposed approach for identification and classification of outbreak patterns.

INTRODUCTION

A foodborne disease outbreak occurs when two or more than two cases of illnesses happen due to same food containing virus, bacteria and toxin in it. A lot of people eat different types of food from different places in a day i.e. office, school, restaurant, home and many others. Many diseases are occurring due to contaminated or poisoned nature, which is very common in some foods at several places and it may cause of death for some individuals. Mostly, some people were not conscious about the food ingredients and also not aware of infectious agent contains in it due to which they get ill or hospitalized. The outbreaks due to several foods occurs which cause of death of some individual from many years.

According to Centers for Disease Control & Prevention (CDC) in United Stated it is estimated that approximately 48 million people (1 out of 6 individuals) get ill, 128,000 were hospitalized and 3000 people were died due to foodborne disease [1]. The investigation of this purpose allows food industry, health officials and agencies to determine the cause of outbreaks and how the food becomes poisoned or contaminated. The analysis of foodborne outbreaks can used to analyze the food inspection authorities to detect the

contaminated food to control the illness. It is very significant to identify the causes of diseases and illness to improve the health impact in the civilians for any country so that the patients and deaths could be minimized.

The Health departments have the responsibility to do the following things to prevent such disease in future as:

- Identify the outbreaks
- Find the germs which cause people to become sick
- Find out the source of outbreaks e.g. contaminated or poisoned food items
- Control the illness to spread
- Prevent the future illness

2. Literature Review

In United States there were estimated 525000 illnesses, 2900 hospitalized and 82 deaths of individuals were happened due to the consumption of pork meal [2]. The analysis of patients was done in the Centers for Disease Control and Prevention (CDC) on the dataset of foodborne disease outbreak in the period 1998-2005. There were mainly 288 outbreaks were recognized due the food prepared by pork meal. The result shows 6372 diseases, 443 hospitalized and 04 deaths happens due to these outbreaks."Staphylococcus aureus toxin" with the ratio 19% in the period 1998-2001 which was shifted to "Salmonella toxin" with the 46% ratio in the period 2012-2015. In resultant, there were16.5 average number of outbreaks was found per year in the period 1998-2015 having range from 10 to 25 and the average number of illness per outbreak was 12 having range from 2 to 333 [3-5].

In Barbados there were 24 foodborne outbreaks were found during the period 1998-2009 having 215 cases of individual illnesses, one hospitalized and no death [6].The dataset in this research was taken of Barbados for the period of 1998-2009. The purpose of this research was to found most frequent etiology causes, food types, ultimate seasons and locations in the Barbados. During this research 37.5% outbreaks were found, which were related to food prepared in the hostel and resorts. The most common agent was "Salmonella Enteritidis phase type 8"occupied in the eggs and other poultry things. The analysis result shows that contamination occurred due to improper cleanliness in the food. So, the proper hygiene and better production practices are required to avoid such outbreaks [7].

In Brazil 30 outbreaks were found due to which 2926 illnesses, 347 hospitalized and no death happened. Some of etiology agents were detected in which most common bacterial pathogens were Salmonella with 30% outbreaks, Staphylococcus aureus with 23.3% outbreaks, Escherichia coli with 10%, Bacillus cereus with 6.6% and Thermotolerant coliforms with 3.3% outbreaks were found during the analysis of data from 2008 to 2014. These agents were occupied in the fruits and vegetables as salads, vegetable salads, caesar salad, tropical salad and raw/cooked salads of cabbage and tomato [8].

A research has been done on the Dutch Salmonella Thompson 2012 outbreak data in which the analysis has been done on the four food products as Minced meat, readymade raw vegetables, ice-cream and smoked fish. The analysis in this research has been also done with "Standard Frequentist method" and "Lasso logistic regression" but among all the "Bayesian analysis" gives better results in identification of mainly etiology agent in the food products. The Bayesian odds ratios of the food products which are not poisoned or contaminated were constantly smaller than the ratio of other food products which are poisoned or contaminated. The analysis has been done by adding missing data in the existing dataset to compare the odds ratios results. The result shows that the model gives similar results for ice-cream, lower odds ratio for minced meat and readymade vegetables and higher odds ratio for smoked fish [9-11].

A nationwide phone survey was conducted during foodbook study from 11139 individuals in Canada to gather the data on consumption patterns of food with 3 and 7 days evoke period [12]. The purpose of this research is to

investigate and identify the source of disease quickly. The analysis was done by using Binomial distribution by calculation the probability of 3 days and 7 days exposure period. The values of 2 days recall period was compared with 3 and 7 days. In results, the major food products don't show any notable difference in this comparison but a pattern is identify that "Salmonella Infantis" was the source of outbreak founded in the chicken mostly in the 3 days recall period [13-14].

In United States approximately 260,000 individuals got ill from contaminated or poisoned fish. The analysis of patients was done in the Centers for Disease Control and Prevention (CDC) on the dataset of Foodborne disease outbreaks due to the consumption of fish in the period 1998-2005. There were mainly 857 outbreaks were recognized due the food prepared by fish. The result shows 4815 diseases, 359 hospitalized and 04 deaths happens due to these outbreaks. "Scombrotoxin" with 34%, "Salmonella" with 26% and "Ciguatoxin" with 23% agents were most common among all outbreaks. Most individuals were hospitalized with "Salmonella" with 31% and "Ciguatoxin" with 23%. The outbreaks in most common types of fish are "Tuna" with 37%, "MahiMahi" with 10% and "Grouper" with 9%. There were 720 diseases happened due to "Scombrotoxin" present in the "Tuna fish" and 660 diseases happened due to "Salmonella" present in the "Tuna fish". The fish prepared in restaurant have 52% and fish prepared in private home have 33% outbreaks [15].

3. Methodology:

A. Dataset

The dataset is "Foodborne disease outbreaks, 1998-2015" of USA with 12 attributes and 19119 numbers of records. The data has been collected from all 17 states of USA. The attributes represents the years, months, states, location (where the food prepared), food, ingredients, species (etiology/agent), serotype/genotype (virus), status (source of illness is confirmed or suspected), illness, hospitalized, fatalities (no. of deaths). The source of the dataset is "Kaggle".

B.

C. Analysis of data (before preprocessing)

The dataset contains missing values in many attributes. By analyzing the dataset it shows the major attributes containing missing values are ingredients and serotype/genotype with 90.19% and 79.56% respectively.

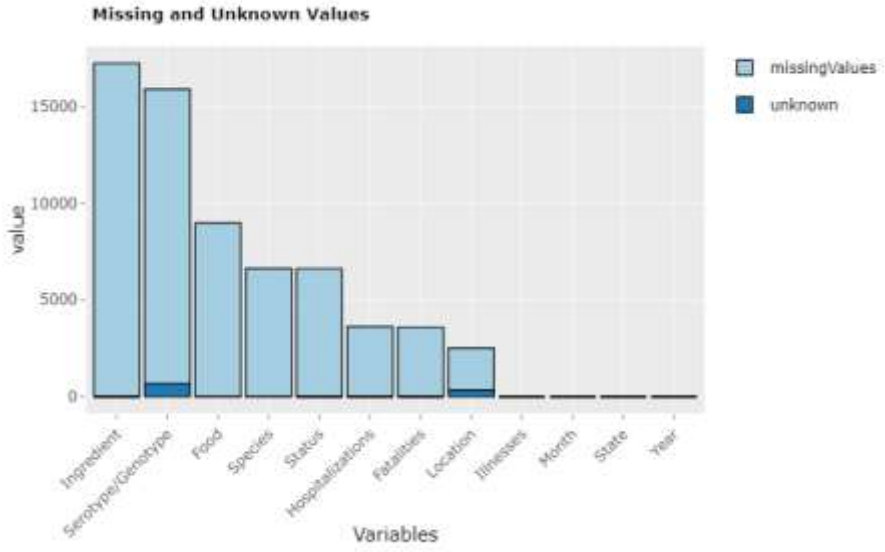

Fig. 1: Missing and unknown values in dataset

Some attributes contains too many distinct values in it mostly in the food attribute as it have 3128 distinct values. So it is difficult to identify the outbreaks in specific food item.

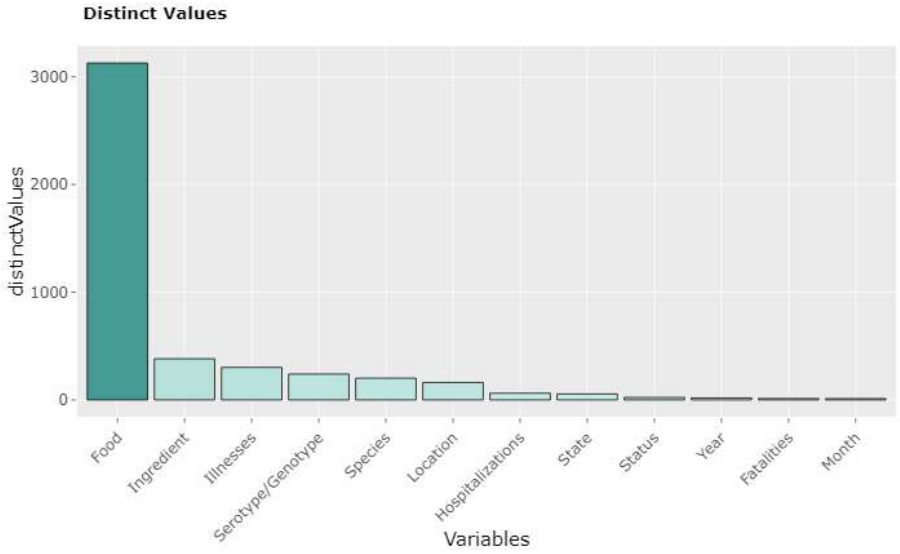

Fig. 2: Distinct values in attributes (before preprocessing)

Attribute	No. of distinct values	No. of Missing Values	Percentage of Missing Values
Year	18	0	0
Month	12	0	0
State	55	0	0
Location	162	2166	11
Food	3128	8963	47
Ingredient	382	17243	90
Species	202	6619	35
Serotype / Genotype	240	15212	80
Status	7	7142	37
Illnesses	302	0	0
Hospitalizations	62	3625	19
Fatalities	13	3601	19

Table1: Analysis of dataset

(Before Preprocessing)

The following issues were needed to resolve before finding pattern and outbreaks as:

- Missing values in attributes
- More than 75% missing values in ingredients and serotype/genotype attributes
- Duplicate records in the dataset
- Too many distinct values in attributes

D. Data preprocessing

The data has been processed to solve the issues which were identified in such a way that:

- All missing values of numeric attributes have been filled with mean.
- The missing value of characters attributes filled with the mode.
- The attributes of ingredients and serotype/genotype attributes have been removed because it has more than 75% missing values in it and by filling it with mode it gives

biased results, so these attributes has been removed.

- The duplicate records in the dataset have been removed.
- Some attributes contains too many distinct values due which the analysis becomes difficult so the attributes has been normalized.

Analysis of data (after preprocessing)

The missing values have been removed after preprocessing the data. The dataset after preprocessing have 10 attributes as the ingredients and serotype/genotype have been discard because it has more than 75% missing values in it. The dataset have 18634 numbers of records in it after removing duplicates.

Attribute	No. of distinct values	No. of Missing Values	Percentage of Missing Values
Year	18	0	0
Month	12	0	0
State	55	0	0
Location	19	0	0
Food	205	0	0
Species	45	0	0
Status	2	0	0
Illnesses	292	0	0
Hospitalizations	52	0	0
Fatalities	9	0	0

Table 2: Analysis of dataset (After Preprocessing)

E. Handling too many distinct values:

There were too many distinct values present in location, food, species and status attributes. There were some values present in these attributes which have their count less than 10 and many attributes contains multiple values in it e.g. Tuna, Seabass, Fin Fish, MahiMahi, Salmon and other types are related to the fish category. So, all types of fish have been normalized into 1 major category named as Fish. Similarly, all sub-categories of food types are normalized into their major categories. The multiple values in records are normalized into single values. The normalization is done to get the better patterns in the dataset.

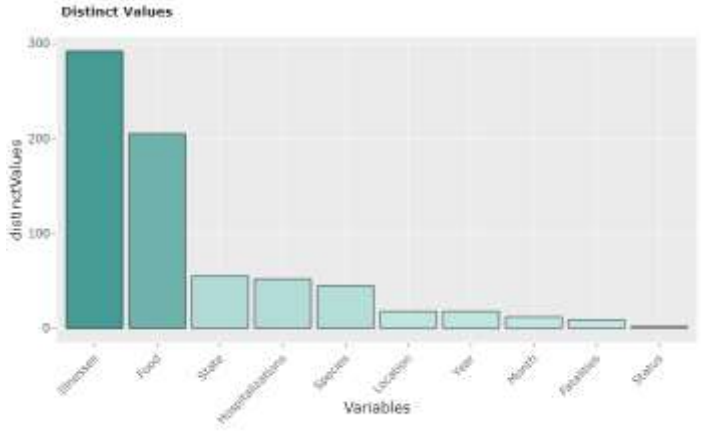

Fig. 3: Distinct values in dataset (after preprocessing)

So, the distinct values have been reduced as Food have 205 instead of 3128 distinct values, Location have 19 instead of 162 distinct values, Species have 45 instead of 202 distinct values and status have 2 instead of 7 distinct values.

F. Tools

The tools used for the analysis are:

- R studio
- Weka

Analysis of outbreaks on locations:

The location wise outbreaks are identified in which it shows the location where most of the outbreaks occurred. The most occurring outbreaks are in the food prepared in the "Restaurant" having 13627 outbreaks due to which 208209 illnesses, 10713, hospitalization and 138 fatalities happened and other locations outbreaks are shown in figure 4.

Location	outbreaks	illnesses	hospitalizations	fatalities
Restaurant	13627	208209	10713	138
Home	2171	35977	3120	88
Ice Cream Shop	1239	41813	739	4
Educational Institutes	358	20082	306	0
Banquet Facility	367	12272	105	1
Fast Food Restaurant	434	6371	584	3

Fig.4: Analysis of outbreaks on locations

Analysis of outbreaks on years:

The year wise outbreaks are identified by which it shows that the number of outbreaks decreased with the passage of time due to which the number of illnesses, hospitalizations and fatalities decreases. In 1998 total number of outbreaks are 1316 identified due to which 27055 illnesses, 1209 hospitalizations and 12 fatalities happened and in 2015 the total numbers of outbreaks are 896 due to which 14111 illnesses, 732 hospitalizations and 8 fatalities were happened.

Year	outbreaks	illnesses	hospitalizations	fatalities
1998	1316	27055	1209	12
1999	1337	24699	1074	10
2000	1405	26033	1263	22
2001	1248	25192	1085	11
2002	1320	24939	1127	14
2003	1089	23079	990	24
2004	1327	28605	905	22
2005	959	19761	758	8
2006	1251	24859	1162	5
2007	1087	20569	905	15
2008	1027	20875	898	11
2009	669	13813	586	7
2010	852	13954	696	20
2011	795	14131	832	12
2012	833	14995	881	20
2013	823	12797	870	14
2014	872	13295	733	23
2015	896	14111	732	8

Fig .5: Analysis of outbreaks on years

Analysis of outbreaks on foods:

The food wise outbreaks are identified in which it shows the most food causing outbreaks in figure 6. The most outbreaks occurred due to the Salad having 1083 outbreaks due to which 28963 illnesses, 853 hospitalization and 12 fatalities happened. There are also more than 300 outbreaks occurred among some foods as 1077 in chicken, 938 in Beef, 824 in Fish, 389 in Pork and 436 in Ice cream.

Food	outbreaks	illnesses	hospitalizations	fatalities
Chicken	1077	18199	1006	8
Salad	1083	28963	853	12
Beef	938	18579	1222	18
Fish	824	5131	518	3
Potato	81	1517	53	1
Pork	389	8639	642	5
Seafood	39	455	29	1
Meat	107	3060	153	18
Cheese	217	5438	457	17
Oysters	34	469	14	0
Milk	214	3722	256	6
Fruits	141	5755	293	12
Eggs	212	3581	336	3
Ice Cream	436	7284	323	16

Fig. 6: Analysis of outbreaks on foods

Analysis of outbreaks on agents having confirmed status:

The analysis of species has been done to determine that the agents are confirmed in the food due to which illnesses occurred. So, it is determined that the following causative agents shown in the figure 7 with their total number of count and they found confirmed. The result shows that the most occurring agent is "Norovirus" with most numbers of confirmed statuses in the food. As total number of counts of Norovirus was 11980 in which 9777 are confirmed.

	confirmed	total
norovirus	9777	11980
salmonella	2189	2396
clostridium	652	1030
escherichia	496	542
campylobacter	349	436
scombroid	319	389
multiple	259	299
staphylococcus	205	550
bacillus	164	386
shigella	152	176
hepatitis	87	88
vibrio	78	135
listeria	56	57
histamine	37	44
toxin	37	156
cyclospora	28	31
cryptosporidium	24	28
mycotoxins	24	30

Fig. 7: Analysis of outbreaks on agents having confirmed status

Outbreaks in Species (causative agents):

The outbreaks of causative agents are identified in which the species having more outbreaks are shown in the figure 8. Among all, "Norovirus" have the most outbreaks as 11980 due to which 216736 illnesses, 4882 hospitalized and 39 deaths are happened.

Species	outbreaks	illnesses	hospitalizations	fatalities
Norovirus	11980	216736	4882	39
Salmonella enterica	2396	57748	6506	66
Scombroid toxin	389	1459	121	0
Clostridium	1030	32830	503	25
Staphylococcus aureus	550	8522	586	3
Bacillus cereus	386	3788	119	0
Campylobacter	436	6380	361	1
Escherichia	542	11797	1934	30

Fig. 8: Outbreaks in Species (causative agents)

The food prepared at "Restaurant" has more outbreaks so, the analysis is done to determine the most frequent agents present in food prepared in restaurant. After analysis it was determine that "Norovirus" is most frequently occurred in the restaurant as shown in the figure 9.

location	species	occurence
restaurant	norovirus	8853
restaurant	salmonella enterica	1364
restaurant	clostridium	585

Fig. 9: Occurrence of agents in restaurant

G. Data Division

The following division of dataset has been done for classification:

- 70% Train set
- 30% Test set

H. Classification methods

The following classifiers are used for classification purpose:

- Decision Tree
- Naïve Bayes
- Random Forest

I. Experiments and Results

1. Decision Tree:

Decision tree is very efficient and powerful learning algorithm used for classification and prediction. It is like a flowchart in which each node represents the attributes, the branch nodes represent the alternate choice between the number and leaf nodes represent the classes.

For this dataset the attribute "species" has been chosen as root node because it has the highest information gain value and the leaf nodes represent the class values "confirmed" and "suspected". The tree is built in such a way to determine that the causative agents found in

the food are confirmed or suspected in the food which cause of illnesses, hospitalizations and deaths. The tree has total 941 numbers of leaves and the size of the tree is 982. The data division for classification is split as 70% is training set and remaining 30% is used as test set.

Result:
As the algorithm is test on 30% set of the whole data having 5732 number of records in which the algorithm classify 4665 correct instances having accuracy 81.38% and the remaining 1067 instances were incorrectly classified having 18.62%.

Confusion Matrix:

	Confirmed	Suspected
Confirmed	4324	114
Suspected	953	341

Table 3: confusion matrix of decision tree results

1. **Naïve Bayes:**
Naïve Bayes classifier uses the probabilistic and statistic approach to classify and prediction based on the prior probabilities. This classifier considers each feature as independent to the other features. The Naïve Bayes classifier considers each feature's probability independently with the prior probabilities to classify it to certain class.
For this dataset the attribute "species" has been chosen as the class label to determine that the causative agents found in the food are confirmed or suspected in the food. The data division for classification is split as 70% is training set and remaining 30% is used as test set. The Naïve Bayes algorithm calculates the probability of all features independently to determine the causative agent in food is confirmed or suspected.

Result:
As the algorithm is test on 30% set of the whole data having 5732 number of records in which the algorithm classify 4076 correct instances having accuracy 71.1% and the remaining 1656 instances were incorrectly classified having 28.89%.

Confusion Matrix:

	Confirmed	Suspected
Confirmed	3301	1138
Suspected	518	775

Table 4: confusion matrix of Naïve Bayes results

2. **Random Forest:**
Random Forest contains individual decision trees in large amount. Each decision tree individually provides the class prediction. The model predicts the class having more count of predictive class by individual decision trees. As multiple decision trees are grown differently so they learn differently which produce high variance. So bagging is used for this purpose which results the low variance. Bagging uses boost aggregation in which the classifier learn by boost aggregate all the decision trees and average them all which gives better results.
For this dataset the attribute "species" has been chosen as the class label to determine that the causative agents found in the food are confirmed or suspected in the food. The data division for classification is split as 70% is training set and remaining 30% is used as test set. The bagging with 100 iterations gives the following results.

Result:
As the algorithm is test on 30% set of the whole data having 5732 number of records in which the algorithm classify 4706 correct

instances having accuracy 82.1% and the remaining 1026 instances were incorrectly classified having 17.89%.

Confusion Matrix:

	Confirmed	Suspected
Confirmed	4245	292
Suspected	734	461

Table 5: confusion matrix of Random Forest results

I. Comparison

Comparison between Classifiers

Classifier	Accuracy	Error
Random Forest	82.1%	17.89%
Decision Tree	81.38%	18.62%
Naïve Bayes	71.1%	28.89%

Table 6: comparison between classifiers

II. Conclusion

The identification and classification of patterns in Foodborne Disease Outbreaks of 55 U.S states of 18 years (1998 to 2015) has been done to improve the food safety. The focus of this research is to find the most causative agents which become the source of disease. The most number of outbreaks identified in year 2000 due to which 26033 illnesses, 1263 hospitalizations and 22 deaths happened. The most number of deaths occurred in 2003. Restaurant and home are the most frequent places of exposure to poisoned food. In food, chicken and salad are most common items having large number of outbreaks due to which most illnesses were happened. Most number of deaths was happened due to beef and meat. The most common causative agent was Norovirus having highest outbreaks. Salmonella entrica was most dangerous causative agent due to which percentage of hospitalization and death is increases. The most frequent item was meat founded in Salmonella entrica outbreaks. So, it is concluded that the most frequent food item is meat having Salmonella entrica agent due to which most people were died. The dataset is classified on the decision tree, naïve bayes and random forest classifiers with 70% training set and 30% test set. The random forest classifies most number of instances from the test set with 82.1% accuracy.

I. Future work

The focus of this research is on the identification and classification of patterns. Random Forest gives the best results in the classification with 82.1% accuracy. In future, the accuracy of classifier could be improved and we can also develop the predictive model to predict the food from unknown places contains harmful agents causing illnesses, hospitalization and deaths.

References

Aladhadh, M. (2023). A review of modern methods for the detection of foodborne pathogens. *Microorganisms, 11*(5), 1111.

Bhattacharya, S., Wasit, A., Earles, M., Nitin, N., Ma, L., & Yi, J. (2024). Enhancing AI microscopy for foodborne bacterial classification via adversarial domain adaptation across optical and biological variability. *arXiv preprint*. http://arxiv.org/abs/2411.19514

Cardim Falcao, R., Edwards, M. R., Hurst, M., Fraser, E., & Otterstatter, M. (2024). A review on microbiological source attribution methods of human salmonellosis: From subtyping to whole-genome sequencing. *Foodborne Pathogens and Disease, 21*(3), 147–160.

Centers for Disease Control and Prevention. (2024, September 17). About the National Outbreak Reporting System (NORS). *Centers for Disease Control and Prevention*. https://www.cdc.gov/nors/index.html

Gomes, K., Araújo, T., Nogueira, T., Oliveira, C., Silva, R., Oliveira, A., Azevedo, M., Almeida, L., & Castro, I. (2025). Advances in whole genome sequencing for foodborne pathogens: Implications for clinical infectious disease surveillance and public health. *Frontiers in Cellular and Infection Microbiology*. https://doi.org/10.3389/fcimb.2025.1593219

He, D., Sun, J., & You, Y. (2025). Raman spectroscopy powered by machine learning methods for rapid identification of foodborne pathogens. *Food Bioscience, 66*, 106281.

Hu, R., Zhang, D., Tao, D., Hartvigsen, T., Feng, H., & Rundensteiner, E. (2022). TWEET-FID: An annotated dataset for multiple foodborne illness detection tasks. *arXiv preprint*. http://arxiv.org/abs/2205.10726

Jaudou, A., Guyot, S., Roussel, S., Berthelot, P., Mallaret, M., & Bonnet, R. (2023). Combination of whole genome sequencing and supervised machine learning provides unambiguous identification of eae-positive Shiga toxin-producing *Escherichia coli*. *Frontiers in Microbiology, 14*, 1118158.

Karanth, S., Patel, J., Shirmohammadi, A., & Pradhan, A. K. (2023). Machine learning to predict foodborne salmonellosis outbreaks based on genome characteristics and meteorological trends. *Current Research in Food Science, 6*, 100525.

Lipman, D. J., Cherry, J. L., Strain, E., Agarwala, R., & Musser, S. M. (2024). Genomic perspectives on foodborne illness. *Proceedings of the National Academy of Sciences, 121*(46), e2411894121.

Liu, D., Zhang, Y., Chen, M., He, Y., Wu, M., Wang, L., & Zhou, Y. (2024). Epidemiological and whole-genome sequencing analysis of restaurant Salmonella Enteritidis outbreak associated with an infected food handler in Jiangxi Province, China, 2023. *Foodborne Pathogens and Disease*. https://doi.org/10.1089/fpd.2023.0123

Njage, P. M. K., Ekman, S., & Buys, E. M. (2024). Epidemiological data mining for assisting with foodborne outbreak investigation. *Foods, 12*(20), 3825.

Taiwo, O., Kumar, P., Singh, P., & Kumari, A. (2024). Advancements in predictive microbiology: Integrating new technologies for efficient food safety models. *International Journal of Microbiology, 2024*, Article ID 6612162.

Wheeler, N. E., Gardner, P. P., & Barquist, L. (2024). Emerging applications of machine learning in food safety. *Annual Review of Food Science and Technology*. https://doi.org/10.1146/annurev-food-071720-024112

Zhou, L., Song, R., Wang, J., & Yang, Y. (2025). Foodborne event detection based on social media mining: A systematic review. *Foods, 14*(2), 239.