



Foodborne Disease Symptoms, Diagnostics, and Predictions Using Artificial Intelligence-Based Learning Approaches: A Systematic Review

Yogesh Kumar¹ · Inderpreet Kaur² · Shakti Mishra¹

Received: 26 April 2023 / Accepted: 3 August 2023 / Published online: 25 August 2023

© The Author(s) under exclusive licence to International Center for Numerical Methods in Engineering (CIMNE) 2023

Abstract

Food-borne diseases have a high worldwide occurrence, substantially impacting public health and the social economy. Most food-borne diseases are contagious or poisonous and are caused by bacteria, viruses or chemicals that enter the body via contaminated food. The most prevalent harmful bacteria (*Salmonella*, *Escherichia coli*, *Campylobacter*, *Clostridium* and *Listeria*) and viruses (norovirus) may cause acute poisoning or chronic disorders such as cancer. Thus, the detection of pathogenic organisms is crucial for the safety of food. Artificial intelligence has recently been an effective technique for predicting pathogens spreading food-borne diseases. This study compares and contrasts the accuracy of many popular methods for making predictions about the pathogens in food-borne diseases, including decision trees, random forests, k-Nearest Neighbors, stochastic gradient descent and extremely randomized trees, along with an ensemble model incorporating all of these approaches. In addition, principal component analysis and scaling methods were used to normalize and rescale the values of the target variable in order to increase the prediction rate. The performance of classification systems has been examined using precision, accuracy, recall, F1-score and root mean square error (RMSE). The experimental results demonstrate that the suggested new ensemble model beat all other classifiers and achieved the average highest 97.26% accuracy, 0.22 RMSE value, 97.77% recall, 97.66% precision and 98.44% F1-Score. This research investigates the predictability of pathogens in food-borne diseases using ensemble learning techniques.

1 Introduction

Food-borne illnesses (FBDs) are a critical and increasing public health issue that causes significant morbidity and mortality worldwide. After ingesting contaminated foods or beverages, various diseases, parasites or microbes cause them to show gastrointestinal symptoms [1]. More than 200 illnesses may spread if people eat food tainted with germs,

viruses, parasites or chemicals like heavy metals. The most severe occurrences often affect the very young, the elderly, those with compromised immune systems and healthy people exposed to very high doses of an organism. In addition, food or water contamination can potentially cause various diseases [2]. Every age group is susceptible to food-borne diseases, which range from cancer to diarrhea [3, 4]. Different forms of FBD have similar symptoms, making it challenging to get a correct diagnosis. In addition to endangering the general public's health, FBD may cause considerable economic losses because of lost productivity, medical costs, hospital stays, epidemiological research costs, and harm to the travel and food industries. Low- and middle-income nations and young children are more prone to getting food-borne diseases than the general population [5]. Most often, contaminated foods are used as a vehicle for the movement of bacteria into the body and other organs, leading to the transmission of food-borne illnesses. Therefore, it is crucial to research the microorganisms responsible for food-borne illnesses. It might be difficult to intuitively classify microorganisms from patient data and disease descriptions

✉ Shakti Mishra
prof.shaktimishra@gmail.com; shakti.mishra@sot.pdpu.ac.in

Yogesh Kumar
yogesh.arora10744@gmail.com;
yogesh.kumar@sot.pdpu.ac.in

Inderpreet Kaur
cecm.bca.ik@gmail.com

¹ Department of CSE, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

² Department of Computer Applications, Chandigarh Group of Colleges, Landran, Mohali, India

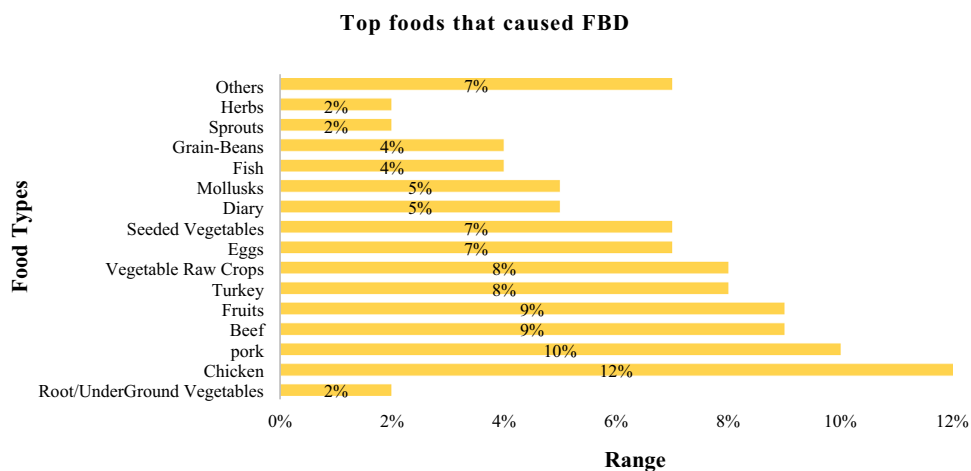
alone [6] because of the variability in clinical manifestations of food-borne disorders caused by various infections. Therefore, research on the microorganisms that cause food-borne illnesses is critical. However, many bacteria that cause food-borne illnesses have similar clinical features, making it intuitively challenging to identify pathogens from patient records and disease descriptions. In addition, conventional pathogen detection methods that rely on lab tests are often time-consuming. Researchers have recently proposed several approaches, notably nucleic acid, biosensor and immunological technologies, for rapidly detecting pathogens in food-borne illnesses; however, these techniques need special tools and their practical usefulness is restricted. Therefore, the pathogens of only a tiny fraction of food-borne illnesses have been discovered, which has a significant influence on the diagnosis of food-borne illnesses, may restrict physician's ability to treat diseases brought on by pathogens and may lead to misdiagnosis. Additionally, the low percentage of recognized food-borne pathogens leads to inadequate data available for analysis, resulting in a negative effect on estimating the disease burden and forecasting outbreaks. The initial manufacturing process carries substantial microbiological hazards. Microorganisms, including human infections, may contaminate various agricultural goods. There are many different types of pathogens and although some may be found in the soil or water, others can be found in animals or people. Figure 1 lists the top 16 items that most often resulted in food-borne diseases [7].

Researchers offer several machine learning-based approaches for diagnosing diseases, predicting disease outbreaks and analyzing the genes of disease pathogens. Food-borne illness research has been enlightened by the effective use of ML in epidemiology; various works have been conducted to tackle foodborne disease issues using machine learning techniques [8]. For example, many classification methods have successfully identified pathogens using images from a near-infrared laser scatter. Compared to conventional

statistical analysis techniques, machine learning techniques quickly provide more accurate results and can manage more extensive and complicated datasets [9]. However, most of this research concentrates on detecting or predicting illnesses, while only a tiny proportion of these studies were conducted to analyze disease pathogens. Consequently, machine learning technologies have become prominent solutions for foodborne disease challenges. The study innovatively uses artificial intelligence (AI) approaches to address foodborne infections. The study endeavors to enhance identifying, diagnosing, and predicting illnesses caused by food AI-based learning techniques, including ensemble learning [10]. This is achieved by analyzing symptoms associated with these diseases. The aforementioned approach can enhance the precision and effectiveness of identifying and forecasting foodborne diseases, thereby assisting in implementing timely intervention and prevention tactics [11]. Implementing artificial intelligence (AI) in this scenario presents a new and encouraging approach towards improving food safety and public health initiatives. A comprehensive investigation is required to understand the impact of data scaling approaches employed in various machine learning models. The significant contribution of the work is summarized as follows:

- Data pre-processing and exploratory data analysis have been performed to clean the data and categories illnesses into state-wise illness, hospital-wise and fatalities.
- Further feature scaling has been performed using principal component analysis to standardize and rescale the target variable values to get a better prediction rate.
- Applied the machine learning models and proposed ensemble learning model, constructed using majority voting to aggregate numerous classification models to predict food-borne disease pathogens.
- Using quality metrics, including precision, accuracy, recall, f-score and RMSE, the applied techniques pre-

Fig. 1 Common causes of food-borne diseases



dicted and evaluate the presence of food-borne pathogens, including norovirus, *Salmonella*, *Campylobacter*, *Clostridium*, *Escherichia coli* (*E. coli*) and *Listeria*.

The following is the rest of the paper: The background investigation review has summarized in Sect. 2, which includes a detailed analysis of previous findings. Section 3 covers the proposed framework, data analysis and pre-processing. Finally, Sect. 4 discusses the various classifiers utilized in the research. The simulation findings and outcomes acquired from performance measure assessment approaches have described in Sect. 5. Finally, the discussion and future improvements are summarized in Sects. 6 and 7, respectively.

1.1 Symptoms and Onsets of Major Foodborne Diseases

Ingestion of a pathogen that subsequently establishes itself (and typically multiplies) in the host's body causes the food-borne disease; similarly, ingestion of a toxigenic pathogen that has established itself in a food item and produced a toxin then causes foodborne sickness in the human host. The two main types of food poisoning are (a) infections and (b) poisonings. Due to the incubation period often associated with foodborne infections, the interval between ingestion and the onset of symptoms is much greater than in the case of foodborne intoxications. There are almost two hundred known food-borne illnesses [12]. The elderly, the young, those with reduced immune function, and healthy individuals exposed to extremely high doses of an organism are more likely to have life-threatening complications. Table 1 displays the most frequent bacteria that cause food poisoning, along with their symptoms, time of onset and prevalence.

1.2 Diagnostics of Foodborne Diseases

Identifying foodborne pathogens linked to numerous obstacles in food has garnered significant interest in scientific

research. The existence of diverse food types, including solid, liquid, meat and ready-to-eat options, poses numerous challenges in sampling, preparation and analysis. Numerous inhibitors in food matrices exhibit considerable efficacy in impeding detection methodologies such as DNA-based assays, PCR techniques and antigen–antibody-specific assays (ELISA). Sample preparation is a challenging task that is carried out before detection. Numerous methodologies have been employed to identify pathogen DNA and toxin in food. However, these assays have encountered setbacks owing to inadequate recovery rate, which is attributed to reduced assay accuracy. For point of care testing (POCT), one of the biggest research issues, several amplification strategies that combine simplicity, cost efficiency, and robustness would be valuable [13]. The procedures for detecting and identifying foodborne pathogens are time-consuming and labor-intensive, necessitating the development of new techniques to detect small concentrations of viable bacterial cells in certain amount of food in near real-time. Recent advances in nucleic acid technology, immunological methods and biosensor design have made it possible to identify even rapidly and accurately trace amounts of live bacteria in food samples. In qualitative and quantitative terms, ELISA is a reliable and accurate method for detecting a wide variety of proteins in a complex matrix. In the food industry, biosensor technology offers encouraging solutions for the portable, rapid and sensitive detection of microorganisms. Figure 2 depicts the various detection strategies for contaminated pathogens, traditional and innovative.

1.2.1 Traditional Methods or Commonly Used Approaches

The conventional techniques employed for microbiological analysis, which involve growing pathogens on media, are known to be dependable and precise. However, these methods are also known to be demanding in terms of time and labor. The methodology entails the amalgamation of the food specimen with a particular enrichment medium, followed by plating onto a suitable media [14, 15]. The

Table 1 Symptoms, onset, and microorganisms/toxins for major foodborne diseases

Toxin/organism	Prevalent symptoms	Approximate symptom onset period (hours)
Norovirus	Vomiting, nausea, diarrhea that is not bloody, and loss of water (dehydration)	12–48 h
<i>Salmonella</i>	Symptoms include fever, abdominal pain, diarrhea, vomiting and headache	6–96 h (often 1–3 days)
<i>Escherichia coli</i>	Symptoms include diarrhea (often bloody), nausea, vomiting, malaise and fever (rare with <i>E. coli</i> O157:H7)	1–10 (typically 3–5) days
<i>Listeria</i>	Symptoms such as high body temperature, low body temperature, pain in joints, weakness, fatigue, sore throat, and enlarged lymph nodes	The duration of time can differ
<i>Campylobacter</i>	The individual presents with symptoms including chills, a high temperature, headaches, arthralgia, prostration, malaise, and swollen lymph nodes	Various durations
<i>Clostridium</i>	Stomach pain, watery stools, putrid stools and sometimes vomiting and nausea	2–36 h (Average: 6–12 h)

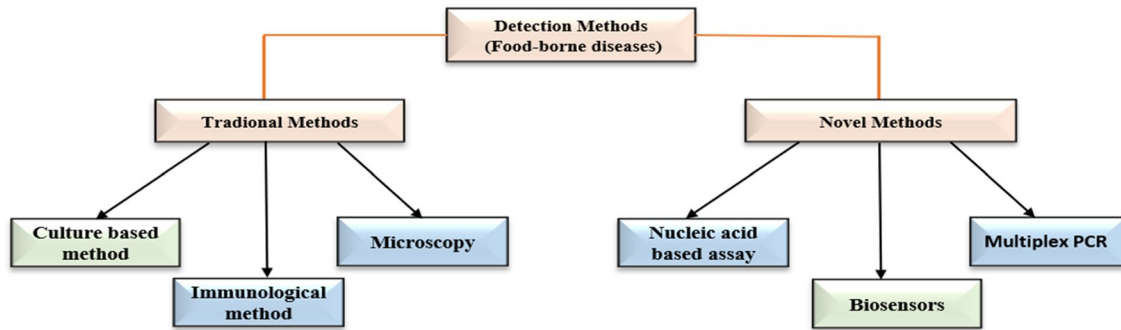


Fig. 2 Various ways for detecting foodborne pathogens

procedure as mentioned above, requires a duration of 48 to 72 h to obtain outcomes and a period of 168 to 240 h for verification. Traditional culture techniques selectively grow just the targeted microbe in a solid or liquid culture medium to limit the development of other microbes present in the food.

1.2.2 Approaches Based on Culture-Based

Pathogens that may be present in food are first cultured in a pre-enrichment medium, then cultured in a selective enrichment medium, and then identified biochemically and confirmed serologically. Notwithstanding, traditional cultural techniques are still advancing and can be amalgamated with other detection techniques to produce more resilient outcomes. There are both quantitative and qualitative ways of studying culture-based methods.

- **Qualitative procedures:** These are employed when the objective is to determine the presence or the absence of an infectious agent in a food sample. In such methods, selective media is utilized to cultivate presumptive colonies from a predetermined quantity of food. The process involves developing uncontaminated cultures, followed by identifying the pathogen through a range of biochemical or serological assays.
- **Quantitative procedures:** The enumeration of microorganisms in food samples by culture method can be achieved by using either the plate counting approach or the most likely number method. Both of these approaches rely on repeated dilution procedures to accurately quantify the microbial load present in the sample. Despite being cost-effective, precise and widely accepted as reliable techniques, the primary drawback of these methods is their protracted analysis duration and demanding manual labor. Typically, the

complete process spans a period of approximately 7 to 10 days [16].

1.2.3 Approaches Based on Microscopy

To ensure the microbiological hygiene of foods and food products, several methods based on the microscopic and optical characteristics of the properly dyed microbe cells have been developed. These methods include:

- The technique of flow cytometry is utilized to quantify the quantity of viable bacteria present in each sample, as well as to examine the viability, metabolic state and antigenic characteristics of bacteria using fluorescent dyes. Using this method, the optical characteristics of cells may be measured when they are individually subjected to a laser beam.
- This method integrates the fundamental concepts of solid phase cytometry (SPC) and Flow Cytometry. Microorganisms are immobilized on a membrane filtering, labelled with fluorescent markers, and quantified through automated enumeration via laser scanning. A computer-controlled moving stage can visually examine individual fluorescent spots by coupling an epifluorescence microscope and a scanning instrument.

1.2.4 Techniques Based on Immunology

Modern immunology serves several functions in medicine, agriculture and other life sciences. Serological detection of illnesses caused by microbial infections has made great strides forward because of the development of the enzyme-linked immuno sorbent assay (ELISA). Antibodies are bound to their target antigens, and the antigen–antibody complex is detected [17]. However, the capacity to identify microorganisms in “real-time” is still lacking in antibody-based detection, which is why ELISA is so famous for developing

pathogenic bacterial and bacterial toxin detection techniques in foods.

These techniques depend on the identification of characteristic stretches of DNA in the chromosomes of the organism of interest (signature sequences). The sequences may be selected to identify a particular microbial genus, species, and even strain. The most frequent and commercially available DNA-based tests for diagnosing foodborne illnesses are probes and nucleic acid amplification methods, however, there are numerous more kinds of DNA-based assays.

- Transducers

Biosensors, which may be broken down into subcategories depending on their transduction techniques, rely heavily on the transducer throughout the detection process. Depending on whether they measure potential, current, conductance or impedance, electrochemical detection techniques are called impedimetric, amperometric, conductimetric or potentiometric. It has been claimed that amperometric detection may be used to identify food-borne diseases such as *E. coli* O157:H7, *Salmonella* and *L. monocytogenes*. Pathogens have been detected using light-addressable potentiometric sensors (LAPS) and immuno ligand assays (ILA). Several types of biosensors based on impedance analysis have been developed for the detection and quantification of food-borne diseases. Research into creating an electronic nose for detecting pathogens has gained traction in recent years.

1.2.5 Novel Methods Used for the Detection of Food-Borne Diseases

Pathogen- and chemical-contaminated foods are both unhealthy and nutritionally deficient. Millions of people all over the globe face serious health risks due to diseases spread via food. A responsive monitoring system requires susceptible, simple, rapid, business-related and transportable detection technologies. These capabilities are shared by sophisticated molecular techniques such as multiplex polymerase chain reaction (PCR) assays, LC-PCR hybridization and many biosensors and electrochemical immunosensors. Together, sensors and the proper signal transduction technology form a biological detecting component that translates the reaction between the goal and the identified element into a meaningful indication. This is where biosensors stand out from the standard method of investigation. These biosensors have several advantages over more traditional methods for detecting pathogens in food and environmental pollutants. As a result, there is a pressing need to enhance methods for rapid, effective and reliable early detection and identification of foodborne pathogens. They are helpful for the rapid diagnosis of food-borne illnesses. More study on the most

significant applications, perception, and pattern, from the analyte to the layout of likely sensors, has been completed to prepare for the detection of potential food pathogens.

1.2.6 Signal-Based Methods: Bioreceptor (Biosensors for the Detection of Food-Borne Pathogens)

An electrical signal may be generated by a biosensor, an analytical instrument that measures biological responses and signals. The system comprises two primary constituents, namely a bioreceptor that facilitates recognition and a transducer that enables the conversion of the recognition event into a measurable and sensitive electrical signal. Antibodies are frequently utilized as bioreceptors and can exist in monoclonal, polyclonal or recombinant forms. To provide enough electron transport to the working electrode, enzymes are selected based on their capacity to bind and their catalytic activity. DNA biosensors now offer intriguing new prospects because to recent developments in nucleic acid identification, particularly the advent of peptide nucleic acid (PNA) [18]. The probe molecule known as PNA exhibits several benefits, including enhanced hybridization properties, the ability to identify single-base mismatches, and increased resistance to chemical and enzymatic degradation. Cell-based bioreceptors (CBBs) are comprised of either complete cells or microorganisms or a distinct cellular constituent, which exhibits the ability to bind to species. Biosensors based on synthetic cells possess a prolonged lifespan and demonstrate greater ease of preservation. Proteins function as a vehicle for the transportation of chemicals, facilitating molecular recognition through various mechanisms. The lectin-based sensor arrays can detect viable cells of both Gram-negative and Gram-positive bacteria, as well as yeast. Moreover, these arrays exhibit the ability to differentiate among five distinct microbial species.

1.2.7 Multiplex PCR Method

About 1.5–2.5 million people each year, mostly small children in developing countries, die from gastroenteritis transmitted from person to person. We have designed, characterized, and applied a multiplex molecular technique to detect *Campylobacter* spp rapidly. In a single examination, this test may identify more than 30 different types of enteropathogenic bacteria from 8 other taxa. We also used this method in the real world to determine the worldwide research standing of *Vibrio cholera* by testing culturally significant pharmaceutical activist samples.

1.2.8 Nucleic Acid-Based Methods

These techniques depend on the identification of characteristic stretches of DNA in the chromosomes of the organism

being studied (signature sequences). The sequences may be selected to identify a particular microbial genus, species, and even strain. There are other more DNA-based test types, however, the most popular and widely used methods for detecting foodborne diseases include nucleic acid amplification methods and probes.

- **Nucleic acid probes:** Probe-based tests are widely used in the food industry due to their convenience. Nucleic acid probes are immobilized to inorganic substrates to prevent their destruction or loss during routine manipulations (such as removing unhybridized DNA by washing). The concept upon which DNA probes rely is relatively simple. It involves hybridizing the DNA sequence of a mystery microbe with a known DNA probe (labelled DNA). When testing for the presence of pathogens in food, microbial cells are lysed to release their DNA, which is then denatured and the probe is applied. The DNA probe is single-stranded and reacts by hybridizing with freed pathogenic DNA in the meal.
- **Polymerase chain reaction (PCR):** Nowadays, polymerase chain reaction (PCR) has replaced culture enrichment as the go-to method for fast exponentially increasing the amount of a particular target sequence. This method is sensitive enough to identify one copy of the target DNA sequence in food contaminated by a single virus. PCR's many advantages over culture and other conventional methods for identifying microbial illnesses include its sensitivity, specificity, precision, rapidity and capacity to detect negligible quantities of the targeted nucleic acid in a sample. There are several different PCR formats available for use in the identification of food-borne diseases.

2 Related Work

Food-borne illness outbreaks are a substantial but avoidable threat to public health, usually resulting in sickness, death, severe economic damage and a breakdown of consumer confidence. As shown in Table 2, we highlighted the significant work that has been done in the domain of food-borne illness detection utilizing machine learning and deep learning techniques. Chenar et al. [19] used a hybrid PCA-ANN modeling technique to create a hybrid model for predicting oyster norovirus outbreaks. This model achieved a sensitivity and specificity of 72.7% and 99.9%, respectively. Zhang et al. [20] built a model for identifying potential outbreaks using data from the CFSA's FDMRS. The eXtreme XGBoost model outperformed all other classification machine learning models, with recall rates and F1-scores of 0.9699 and 0.9582, respectively. Chenar et al. [21] created the ANN-2Day model, which uses artificial intelligence to predict oyster viral outbreaks. The ANN-2Day model's

accuracy score of 99.83 percent demonstrates its usefulness in forecasting the danger of norovirus sickness epidemics to human health. Min et al. [22] investigated a gadget lateral-flow test analyzer that detected *Salmonella* spp. using machine-learning approaches. Images of test lines are used to calculate concentrations Integrating colour spaces with SVM and K-nearest neighbors classification culminated in a high level of accuracy, according to the researchers (95.56 percent). Nguyen et al. [23] utilised a sample dataset of 5278 *Salmonella* genomes to construct XG Boost based machine learning methods for determining MICs for 15 antibiotics. Within a simple twofold dilution step, the MIC forecasting models attained an overall average accuracy of 95%. Without prior information on the bacterium' underlying gene makeup or resistance features, the model calculates MICs. Polat et al. [24] used a 539-point data set to train and analyze classification methods such as artificial neural networks (ANNs), the nearest neighbor method (kNN) and support vector machine (SVM). Generic *E. coli* exhibited the highest accuracy (about 75 percent) of all the algorithms studied, followed by enterococci (65 percent) and total coliforms (60 percent). Additionally, classifiers calculated turbidity accuracy 6–15 percent higher, ranging from 62 to 66 percent. According to Amado et al. [25], this study intends to develop prediction models for identifying microorganisms such as *E. coli* and *Staphylococcus aureus* in raw meat using machine learning approaches. The models perform well, with 94.97, 91.84, 97.57, 61.46 and 66.84% accuracy rates, respectively.

Wang et al. [9] used case data from the National Food-borne Disease Surveillance Reporting System to construct a machine learning-based categorization strategy for bacteria associated with food-borne disorders. Lupolova et al. [26] developed a SVM classifier for predicted proteins based on whole-genome sequencing data. The same approach was utilized to demonstrate the classifier's capacity to foresee a zoonotic threat using correctly identified human Vs. bovine *E. coli* isolates (83% accuracy) and *E. coli* O157. Hieura et al. [27] estimated viable bacterial counts in different meals using an eXtreme gradient boosting tree, a machine learning approach. The mean square error was around 1.0 log CFU between the actual and predicted levels. Njage et al. [28] used Basic Local Alignment to match the assemblies to a collection of 136 virulence and stress resistance genes (BLAST). Due to the lack of notable difference in accuracy amongst the algorithms, we chose an SVM classifier with a linear kernel for further research. Borujeni et al. [29] simulated *Listeria monocytogenes* and *Escherichia coli* bacterial development using sigmoid functions, including Logistic and Gompertz functions, and (LSSVM)-least square support vector machine-based techniques. This study enhanced the parameters of the LSSVM by applying a non-dominated sorting genetic algorithm-II (NSGA-II). Bandoy et al. [30] used GWAS, a machine learning method and bacterial

Table 2 Comparative summary of food borne pathogens

Reference	Type of illness	Datasets	Techniques used	Limitations	Findings
[19]	Norovirus	Satellite remote-sensing data	ANN PCA-ANN (hybrid)	Remote sensing data may not be accessible on cloudy days	PCA-ANN outperformed with 72.7% sensitivity and 99.9% specificity
[20]	Norovirus	Taken from FDMRS database	DT, RF, SVM, AdaBoost, XGBoost, GBDT, NB, LR	Restricted dataset, insufficient to detect early indications of the CFSA platform	XGBoost algorithm has the best recall rate of 0.9699 and F1-score of 0.9582
[21]	Norovirus	Data collected from harvesting areas	ANN, ANN-2Day Model	Model-predicted hazard is not tied to epidemic size or human infection rate	The ANN-2Day model has a specificity of 99.84%, accuracy of 99.83%
[22]	<i>Salmonella</i>	Images were taken by Smartphone	NB, k-NN, DT, SVM and Neural Networks	Unable to discriminate among control and sample approaching detection limit	The accuracy of kNN and SVM classifiers was 95.56%
[23]	<i>Salmonella</i>	Taken from project NARMS	XGBoost	XGBoost can't detect whether a k-mer is linked to resistance to antibiotics	Model has achieved accuracy of 95%
[24]	<i>Salmonella</i>	Data taken from agriculture pond	kNN, ANN, and SVM	Affecting the performance of algorithms using various data sets	k-NN outperformed with the accuracy of 77.15%
[25]	<i>Escherichia coli</i>	From open-source terminal Tera term	ANN, NB, k-NN, SVM, RF	Overfitting problem	RF outperformed with accuracy of 97.57%
[9]	<i>Escherichia COLI</i>	Data collected from NFDARS	RF, DT, GBDT, adaptive boosting techniques	Scope of disease data was limited as its collected from surveillance system	GBDT model achieved the accuracy of 69%
[26]	<i>Escherichia coli</i>	Dataset collected from https://figshare.com/s/7a3ededa8cedd95b9fb7	SVM	Dataset failed to capture all host-specific genetic information, improper collection of metadata	SVM obtained accuracy of 83% in the case of <i>Escherichia coli</i>
[27]	<i>Listeria</i>	Taken from ComeBase (www.combase.cc)	XGBoost	pH & aw values are missing from the data	Model has achieved RMSE score 1.13 and model learning rate is 0.01
[28]	<i>Listeria</i>	NGS data	RF, SVM, GBM, NN and LB	Existing risk assessment attempts are limited by dose-response methods	SVM outperformed with 89% accuracy
[29]	<i>Listeria</i>	Data collected from published study	LSSVM (least square support vector machine), NSGA-II-LSSVM(Hybrid)	Limited dataset, optimal kernel function required	LSSVM clearly outperformed sigmoid functions
[30]	<i>Campylobacter</i>	Re-used wet-lab data	XGBoost	Lack of relevant statistical models in microbiology	XGBoost optimizes the functional space by strengthening weak learners
[31]	<i>Campylobacter</i>	Created a synthetic dataset	DT	Scarcity of data, low quality data	Learning rate=0.5, Se= 1102/(1102+315)=0.78 and Sp= 1 571 036/(1 571 036+ 82 378)= 0.95 are the sensitivity and specificity of model
[32]	<i>Clostridium</i>	Downloaded from FDA	SVM, NB, LR, BERT	Imbalanced dataset	BERT outperformed with recall 0.71, precision 0.78, and F1 score 0.74
[33]	<i>Clostridium</i>	Data collected from previous published study	Boosted gradient classifier	Poor accuracy may be of little benefit in a clinical environment	On balanced sets, boosted gradient classifier scored a median accuracy score of 64%

NB Naïve bayes, kNN k-nearest neighbor, DT decision trees, RF random forest, SVM support vector machine, ANN artificial neural network, XGBoost extreme gradient boosting, LSSVM least square support vector machine

community genetics to establish the infection mechanism. Using 1.2 million SNPs and indels, our method prioritized the relevance of *porA*-related alleles in *Campylobacter jejuni* for 30 years. As a result, an intestinal and extraintestinal group named PathML has differential *porA* mutations that induce abortion. Hill et al. [31] used a food chain approach to estimate the advantages of integrating genetic and epidemiological data. Studying food supply chain dynamics would likely provide additional information on foodborne risk. To automate the identification of complaints of dangerous food goods, Maharana et al. [32] employed ML (machine learning) approaches and over-and under-sampling techniques to corresponding data. The best method for detecting dangerous food reviews was Bidirectional Encoder Representation from Transformations, with an F1 score of 0.74, accuracy of 0.78 and recall of 0.71, respectively. OLM et al. [33] employed the ML approach Boosted gradient to find microbiological characteristics predictive of NEC. Necrotizing enterocolitis (NEC) is a life-threatening intestinal condition that affects preterm babies. Genetics, bacterial strain varieties, eukaryotes, bacteriophages, plasmids and growth rates are all considered.

3 Proposed Design

The datasets, suggested strategy and methodologies employed for early prediction of food-borne pathogens are discussed in this section. As seen in Fig. 3, the proposed framework makes an effort to enhance food-borne pathogens prediction results via ensemble learning techniques. In the proposed methodology, data is collected and prepared from an open access data set, Kaggle, as stated under Sect. 3.2, followed by data standardization and feature scaling techniques, outlined in Sect. 4, before being used in the proposed strategy. Furthermore, the data has been augmented and segregated for validation. The following automated learning algorithms are used to classify the pathogens (*norovirus*, *Salmonella*, *Campylobacter*, *Clostridium*, *Escherichia coli* (*E. coli*) and *Listeria*): The decision tree (DT), the random forest (RF), the k-nearest neighbor (KNN), the stochastic gradient descent (SGD), the extremely randomized trees (Extra Tree), and the Ensemble Technique (Hybridization of the DT + RF + KNN + SGD + Extra Tree). Section 4.4 provides a comprehensive description of all of the ML (Machine Learning) techniques used in the research. Further, the models are tested and compared based on their ability to predict the future using various evaluation criteria (Sect. 5).

3.1 Preprocessing

Data must be pre-processed after collection. As shown in Fig. 2, pre-processing is the practice of cleaning, checking,

and organizing data to create a useful dataset. Before creating the machine learning (ML) model, the most time-consuming operation is data preparation. It raises the dataset's quality, giving machine learning algorithms more data to work with. Additionally, good pre-processing may directly impact the model's ability to produce accurate predictions. In machine learning, data pre-processing is used to clean and prepare data to meet the model's requirements [34]. As discussed in the following sections, we used cutting-edge techniques to cleanse and preprocess the data in our investigation.

3.2 Dataset Description

The Pandas software was used to import the CSV dataset, which can store enormous volumes of data for mathematical computations. This dataset, which can be found at "<https://www.kaggle.com/datasets/cdc/foodborne-diseases> select=outbreaks.csv," provides information on foodborne illnesses reported to the CDC between 1998 and 2015. This dataset contains twelve attributes as shown in Table 3. The dataset was split into two parts for the experiment: a 75% training data set and a 25% testing set. Year, state and the site where the meals were prepared, confirmed food vehicle and infected component, etiology status, cumulative chronic conditions, hospital visits, and deaths are just some of the data fields. Many epidemiological studies fail to identify food vehicles; hence the food vehicle variable is left undefined in these cases. Figure 4 illustrate the disease count reported between 1998 and 2015 outbreak, Norovirus was responsible for the majority of outbreak-related infections (53%), followed by *Campylobacter* (24%) respectively. To complete different tasks, we employed many libraries. Some of the standard libraries are NumPy, Pandas, Seaborn, Scikit—Learn, Matplotlib and mlxtend. Matplotlib is a free and open-source toolkit for producing graphs and charts using numerical computing. It works with the Panda Framework as well as the Numbly Array. Pandas is used to manipulate and analyze data, while Seaborn is the best open-source data visualization tool available. It is faster than Python lists, and it can hold more data for scientific calculations.

Pandas were used to import the dataset, and Scikit-Learn, a free machine learning framework developed in Python, was utilized for the analysis. The dataset has some missing values, which may be imputed or approximated using the existing data, a more effective technique. We imputed missing data using the Scikit-learn Simple Imputer class. On our dataset, we utilized Label Encoder to encode target labels ranging from 0 to n class-1. We may artificially lower the proportion of variance in your dataset by supplying new values close to the mean via data imputation. Missing values may be imputed using a constant value specified by the user or the columns' statistics (mean, median, or most frequent).

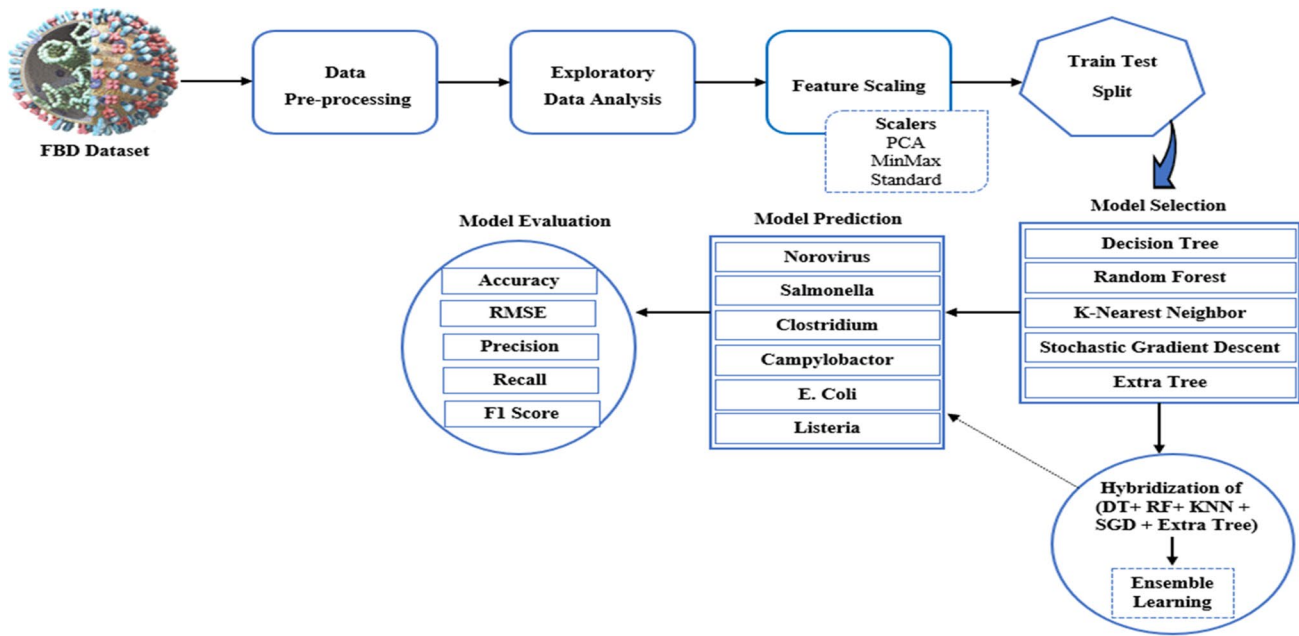


Fig. 3 Framework for predicting pathogens in food-borne diseases

We may artificially lower the proportion of variance in your dataset by supplying new values close to (or equal to) the mean via data imputation.

3.3 Distribution of Illness, Hospitalizations and Fatalities Based on Months, States and Years

The etiology, month, year and state in which the outbreak occurred and the number of illnesses, hospital admissions, and fatalities during the outbreak are shown in graphs. In addition, it allegedly involved food(s), sites where meals were prepared & consumed, and factors that contributed to the outbreak are all included in the outbreak study. As shown in Fig. 5a, May has the month with the most significant number of diseases, more than 1800, followed by December as the month with the following highest significant number.

On the other hand, September has the fewest range of illnesses reported, which is 1250. In addition, according to Fig. 5b, Florida has the highest total number of illness occurrences. According to graph Fig. 5c, the most instances were reported in 2000, while the fewest were 2009, i.e., 600.

In Fig. 6a shows that May has the most instances, with over 1500 patients hospitalized, followed by December with 1500 cases. On the other side, September had the fewest hospitalized cases (1200). Furthermore, California and Florida have the highest number of cases hospitalized due to food-borne infections, with over 2000 instances each, while Guam has the lowest number of cases hospitalized, as seen in Fig. 6b.

Finally, the most significant and lowest numbers of hospitalized patients were recorded in 2004 and 2009, as depicted in Fig. 6c.

May and December seemed to have the highest number of deaths (over 1500), whereas September had the lowest number of fatalities, as seen in Fig. 7a. California had the most significant number of deaths, with 2000, while Florida had the second-biggest number of fatalities, with 2000. According to Figs. 6, 7b, the states with the fewest deaths instances were Washington, DC, and Montana. As seen in Fig. 7c, the years 2004 and 2009 had the most deaths (> 1000), while the year 2009 had the fewest (600).

4 Feature Scaling

4.1 Principal Component Analysis (PCA)

We use feature scaling to normalize each feature’s metric measures to avoid big-number features from overpowering learning. Principal Component Analysis reduces data sets so they may be examined faster. If there are N eigenvectors, the equation shows the explained variance for each Eq. (1).

$$\text{Explained variance} = \frac{\alpha_i}{(\alpha_1 + \alpha_2 + \dots + \alpha_n)} \tag{1}$$

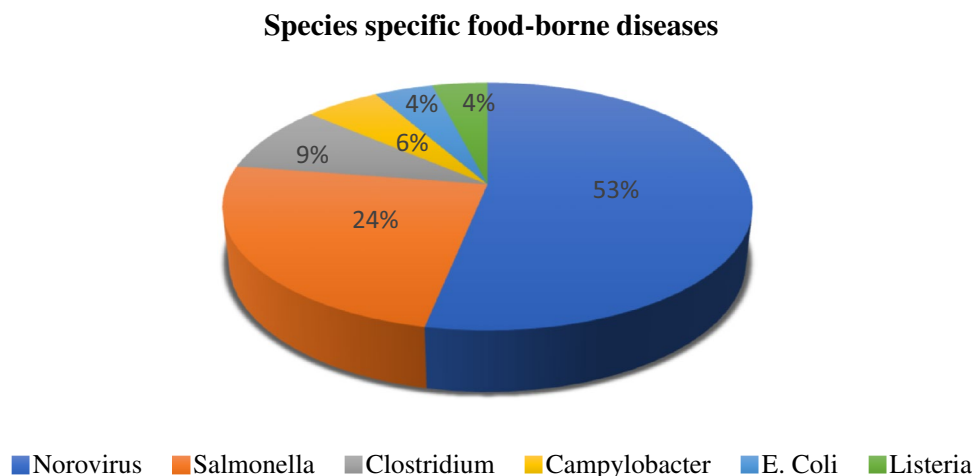
α_i in the equation represent the matching eigenvalue. The standard scalar function is used to standardize data. The distributions of an independent variable are normalized to

Table 3 Description of sample dataset used in study

Year	Month	State	Location	Food	Ingredient	Species	Serotype/Genotype	Status	Illnesses	Hospitalizations	Fatalities
1998	February	Florida	Catering service	Stir-fry, unspecified; caesar salad	-	Norovirus genogroup I	-	S	35	1	0
1998	January	California	Private home/residence	Lasagna, unspecified; eggs, other	-	<i>Salmonella enterica</i>	Enteritidis	S	26	3	0
2015	December	Utah	Private home/residence	Meal replacement powder	-	<i>Salmonella enterica</i>	Virchow	S	3	1	0
2015	December	Ohio	Restaurant	Ethnic style	-	<i>Clostridium perfringens</i>	-	S	2	0	0
1998	January	Florida	Restaurant	Ethnic style, unspecified	-	<i>Clostridium perfringens</i>	-	S	3	0	0
2002	March	Washington	Restaurant	Sandwich, beef	-	<i>Campylobacter unknown</i>	-	S	2	0	0
2011	June	Pennsylvania	Restaurant	Milk, whole milk unpasteurized	Milk, whole milk unpasteurized	<i>Campylobacter unknown, Escherichia coli, Enterotoxigen</i>	-	C	5	0	0
2002	October	Oregon	Catering service	Lasagna, chicken	-	<i>Escherichia coli, Enterotoxigenic</i>	O27	C	49	0	0
2015	August	Washington	Catering service; fair/festival	Milkshake	-	<i>Escherichia coli, Shiga toxin-producing</i>	O157	C	13	5	0
2015	August	Washington	Grocery store	Kale, unspecified; pineapple juice, unpasteurized	-	<i>Escherichia coli, Shiga toxin-producing</i>	O121	C	2	0	0
2014	November	Washington	Hospital	Milkshake	-	<i>Listeria monocytogenes</i>	-	C	2	2	0

S suspected, C confirmed

Fig. 4 Total count of reported food-borne diseases (1998–2015)



have a mean of zero and a variance of one. By generating the relevant statistics on the training set instances, each attribute performs centering and scaling independently to biased outcomes. Making the data identical scales may assist prevent this problem. To achieve this mathematically, remove the mean and divide by the standard deviation to every variable.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \tag{2}$$

After standardization, all parameters will have the same scale. However, since variables are highly connected, they contain duplicate information. To determine these relationships, we generate a covariance matrix. The covariance

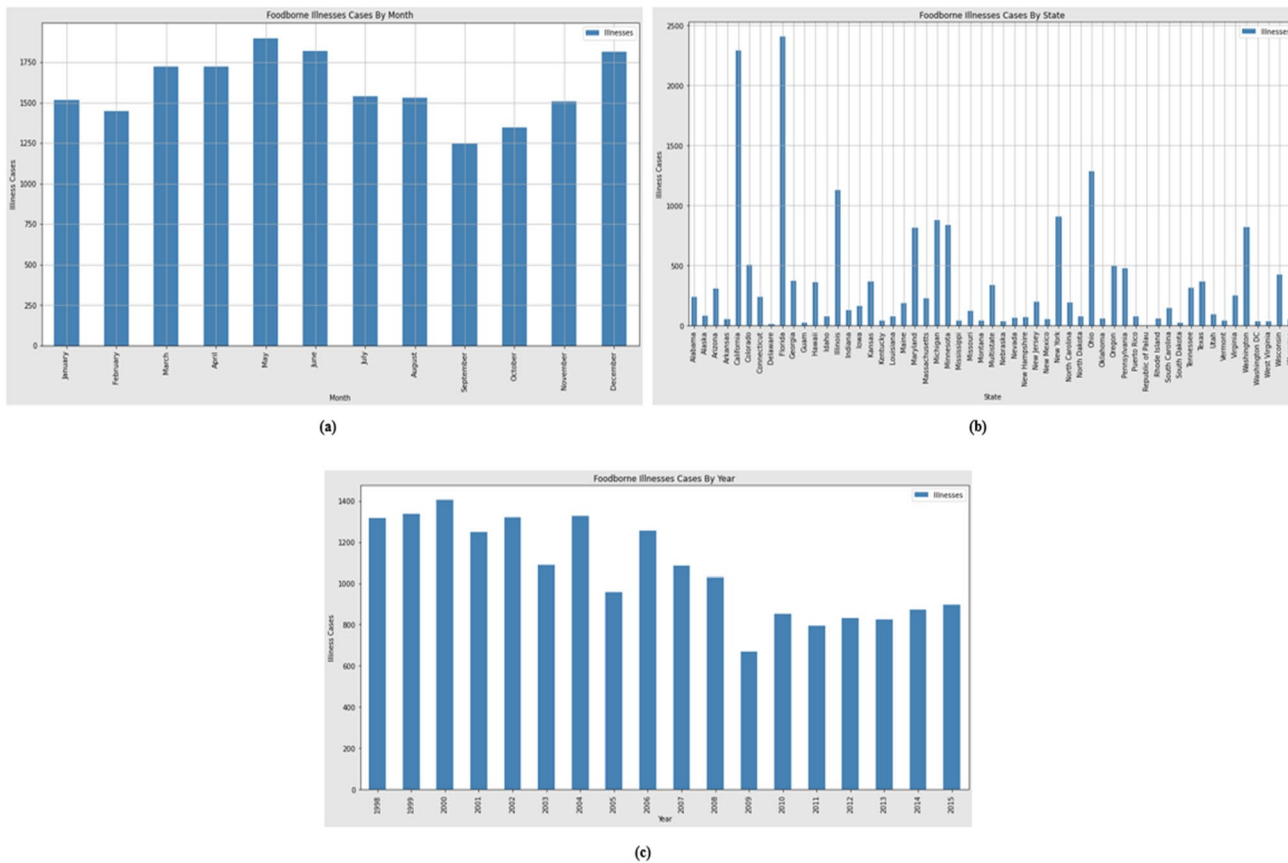


Fig. 5 Illnesses based on months (a), states (b) and years (c)

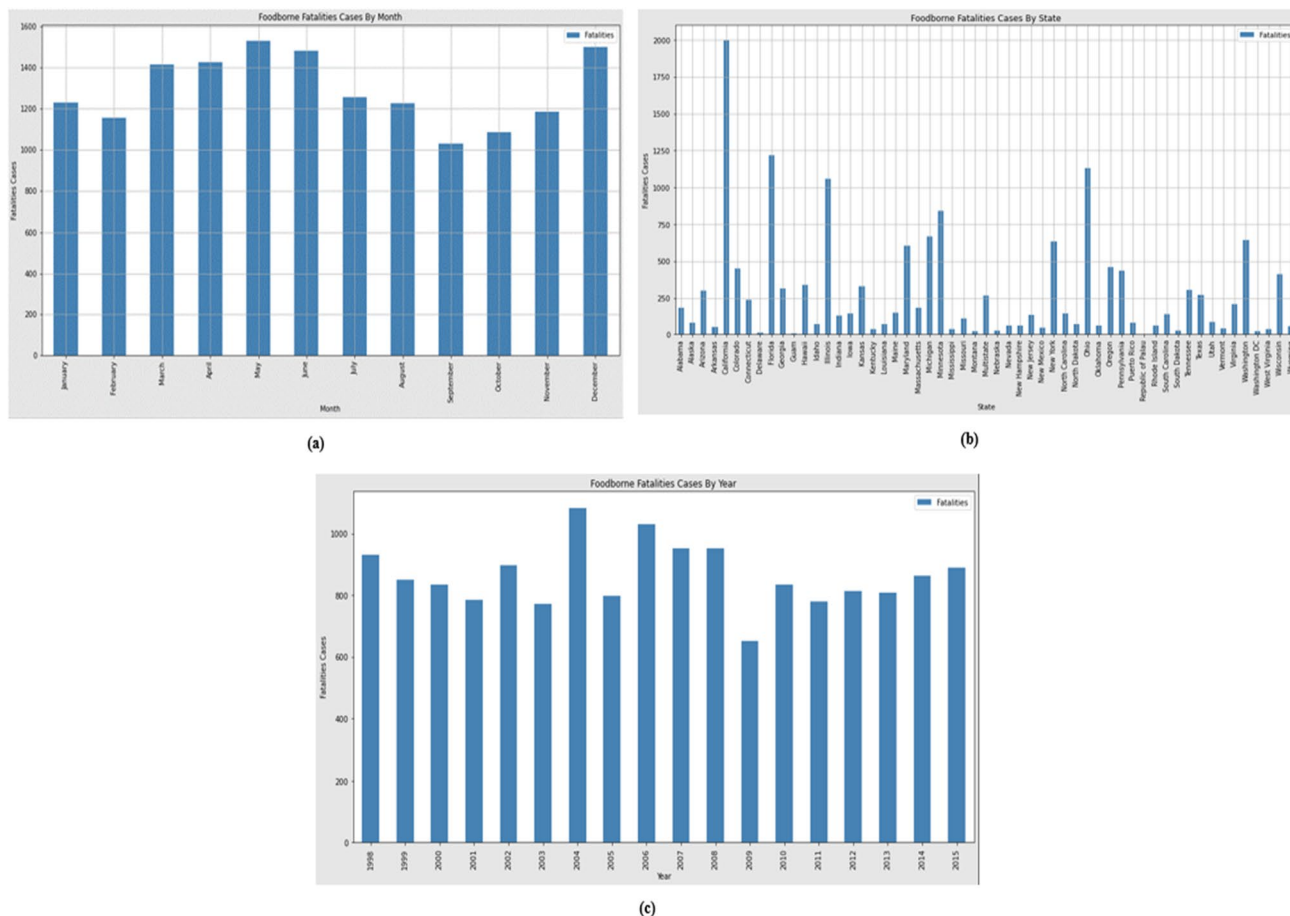


Fig. 6 Hospitalizations based on months (a), states (b) and years (c)

matrix is a p-square matrix that includes the covariances and all potential pairings of the initial variables as entries. To get the data’s principal components, we must first calculate the correlation matrix’s eigenvectors and eigenvalues. New parameters are created by merging or combining previous variables to generate major components. The variables are formed once the data from the actual parameters are condensed into the uncorrelated starting components. It is possible that determining the eigenvectors and sorting them in descending order of their eigenvalues can help us determine the relative importance of the various components. Here, we choose out which of these elements have high eigenvalues and keep the others, or we throw out the ones with low eigenvalues. Therefore, a feature vector is simply a matrix with the eigen-vectors of the elements we desire to retain as columns. In the last step, we use the feature vector computed from the covariance matrix’s eigenvectors to reorient

the data such that it lies along the axes suggested by the critical components. To do this, we multiply the features vector’s transposition by the transposed original data.

$$\text{Final Dataset} = \text{Feature Vector}^T \times \text{Standardized Original Dataset}^T \tag{3}$$

4.2 Scaling Methods (MinMax Scaler and Standard Scaler)

Numerous machine learning methods perform more effectively when numeric input parameters are scaled to a normal distribution. We use feature scaling to normalize metric measures for each feature to avoid learning becoming overwhelmed by large-number features. Normalization or feature scaling, is the last stage in pre-processing; it standardizes

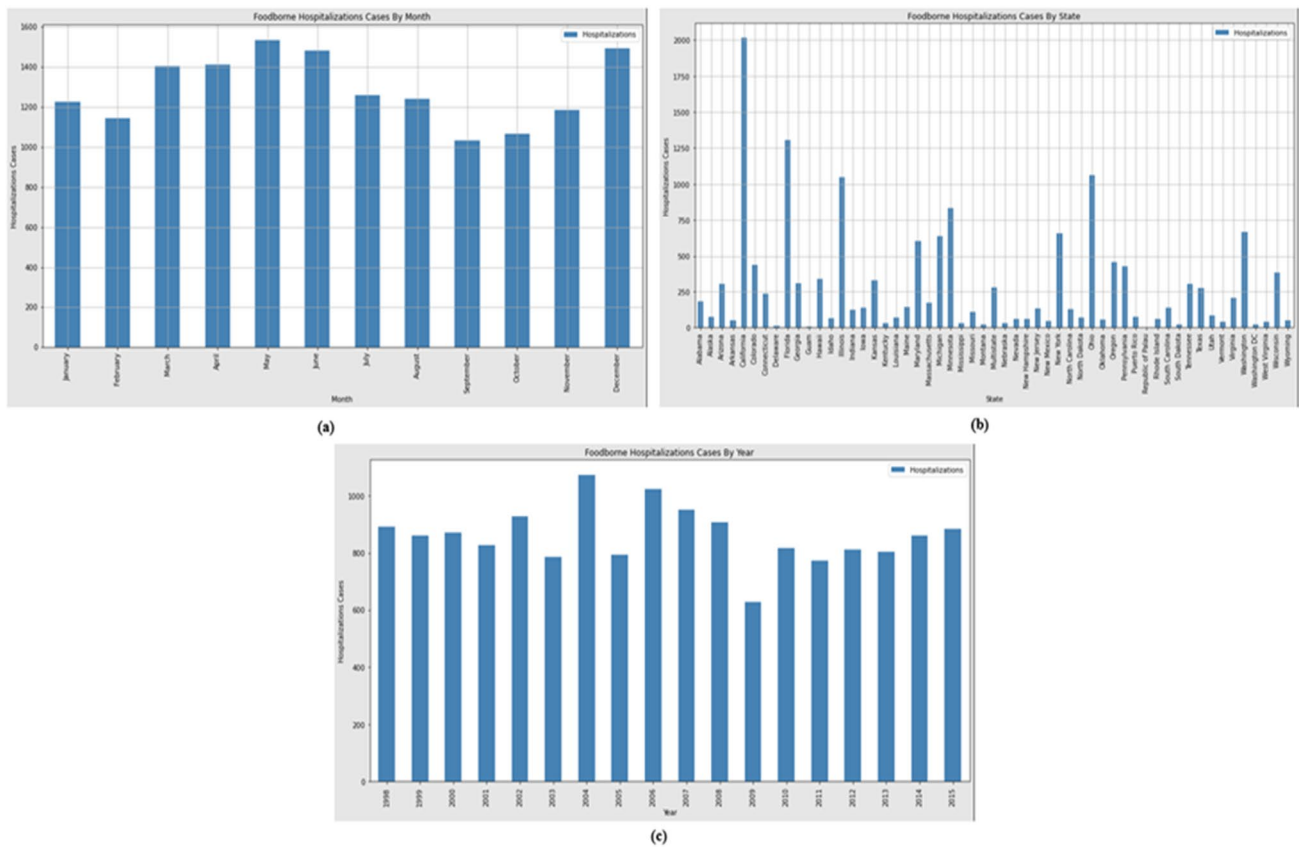


Fig. 7 Fatalities based on months, states and years

data propertie’s range. Most machine learning algorithms function well when numerical input data are scaled to a specific range. Each attribute has a range, such as a year, month or a number of diseases, and if these values are supplied during the training phase, the model will be unable to understand the skewness in the data-set range [35]. It has no concept of years or counting; all it recognizes are numbers that span an extensive range, resulting in an inadequate model. Additionally, feature scaling is required since the size of the data’s input variables may fluctuate. In Python, the sklearn package contains tools for pre-processing raw feature vectors into a format that downstream estimators can comprehend. Minmax scaler and standard scaler are two scaling techniques for continuous variables used in machine learning. The minmax scaler is a particular case of a scaler in which the minimum and maximum values are scaled to 0 and 1, respectively. While the standard scaler adjusts all values within min and max to fit inside a range defined by the min and max values.

$$\frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{4}$$

standard Scaler is based on the standard normal distribution, implying that the data included inside each feature follows a normal distribution [36]. The scaling produces a distribution with a mean of 1 and a standard deviation of 1 due to the transformation. The following formula is used to measure the standard scaler score for a data sample x: Where sd indicates the standard deviation of x

$$\frac{x_i - \text{mean}(x)}{\text{sd}(x)} \tag{5}$$

4.3 Dataset Split

Models are trained first, and then tested against test data [37]. The dataset is then randomly split into train and test groups for this purpose (three-fourths of data cases). Subsequently, we reported the outcomes associated with the testing data.

4.4 Model Selection and Hyper-Parameter Tuning

The methodology and hyper-parameter values utilized may significantly influence the model’s performance, but

choosing them requires much experience and many manual iterations. Computer science academics have developed several automated selection approaches for algorithms and hyper-parameter values for a particular supervised machine learning issue to make machine learning accessible to layperson users with low computing experience. The regularisation hyperparameter controls the model's capacity by determining the model's adaptability. In the case of overfitting, the model loses some of its ability to predict novel test data because it is too malleable and adjusts too much during training. Effective model capacity control prevents overfitting. Therefore, the hyper-parameters must be suitably adjusted. This section summarizes the categorization methods used in the research to predict various food-borne illnesses.

4.4.1 Decision Tree (DT)

To examine rules, Decision Tree picks target variables and branches to show as a hierarchical structure. The hierarchical link between choice variables and target variables may be discovered and used for prediction using a trimmed DT [38]. A DT is a multivariate analytic approach that uses supervised machine learning. In Machine Learning, the Decision Tree technique may be used to tackle both classification and regression issues [39]. CART or Classification and Regression Trees, is another name for it. Decision trees are divided into two categories. They are classified according to the sort of target variable they possess. The decision tree is referred to as a 'categorical variable decision tree' if the target variable is categorical. A 'continuous variable decision tree,' on the other hand, it has constructed training models for a continuous objective variable that may be used to anticipate the target class, and decision criteria are based on past data (training data). The node(root) of the tree has been the most distinguishing attribute. These trees are generated utilizing rules-based upon "if-then" classifiers, with different rules for each edge between root and leaf. "gini" is the criterion used to separate nodes. The min sample split is set as "10" with max depth of the tree as 5 in this study.

4.4.2 Random Forest (RF)

A decision tree-based ensemble model called random forest can address the issue of decision tree's limited generalizability [40]. It constructs several decision trees to arrive at the outcome and employs voting procedures. Each tree obtains its training data via a different proportional sampling of its features using a replacement sampling technique. The RF classifier uses k decision tree classifiers,

each of which is repeatedly applied to the decision function to classify all of the labels in the data. These are built on several decision trees and work on the ensemble learning principle by using a bootstrap aggregating (bagging) approach to their trees during the training phase [41]. The results are counted similarly to how majority voting is conducted. Bootstrap-True, Max depth-00, Max features-3, Min sample leaf-5, Min sample split-12, N estimators = 1000 are the parameters used to generate the Random Forest classification model.

4.4.3 K-Nearest Neighbours (KNN)

The K-nearest neighbor (KNN) technique is an essential machine learning algorithm that classifies data points by computing their distances. This approach is widely used in statistical estimates and pattern identification. KNN keeps track of its examples and classifies the most recent ones using a similarity metric [42]. It will not make any previous assumptions as it is a non-parametric method [43]. The Euclidean Distance is used by the KNN method to determine the similarity between data points. A case classification will be determined by a majority of votes of its neighbors, with the case being allocated to the most frequent class among its K closest neighbors as determined by a distance role. The KNN algorithm was chosen because it allows us to classify data using the most significant number of nearby values. If many similar data were examined, the risk of generalizing based on a small number of data would be reduced, and the risk of misclassifying non-diseased people as diseases people would be reduced. Sampling approaches and hyper-parameter tweaking are used to fine-tune the model's performance. Each test sample's predicted class in KNN is assigned to the class that most of its k -nearest neighbors in the training set belong. Assume the training set exists $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, x_i is the instance's feature vector and $y_i \in \{c_1, c_2, \dots, c_m\}$ is the instance's class, $i = (1, 2, \dots, n)$, the class y of a test instance x may be represented by:

$$y = \operatorname{argmax}_{c_j} \sum_{x_i \in N_k(x)} 1(y_i = c_j), \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m \quad (6)$$

in which $I(x)$ is an indicator function, with $I = 1$ when $y_i = c_j$ and $I = 0$ otherwise; $N_k(x)$ is the field involving x 's k -nearest neighbours. Weights = uniform neighbours = 100 are the parameters used to build the KNN classification model in this work.

4.4.4 Stochastic Gradient Descent (SGD)

A popular machine learning (ML) method for model optimization is the SGD. It could make it possible for

linear classifiers using convex loss functions, including neural nets [44], SVM classifiers and LR, to learn discriminatively. SGD is an improved gradient descent-based approach. Because it employs an estimated gradient rather than an actual gradient via randomly subsampling a whole training dataset, this approach is considered a stochastic approximation of gradient descent optimization [45]. The technique is popular for datasets containing redundant samples due to its excellent efficiency and ease of implementation. SGD, on the other hand, is seldom used in landslide susceptibility evaluations, which need to be thoroughly investigated. The parameters used for building the SGD classification model are the loss function used to construct the model is the hinge, Penalty = 12, Max_iterations = 500.

4.4.5 Extremely Randomised Tree

The Extra-Trees algorithm creates a top-down ensemble of unpruned decision trees. Unlike previous tree-based ensemble approaches, it separates nodes randomly and grows trees from whole learning samples instead of a bootstrap replica [46]. These parameters are K , the number of randomly picked characteristics per node, and n_{\min} , the minimum sample size for dividing a node. It is employed numerous times with the original learning set to produce an ensemble model (M trees). The final forecast is based on the majority vote in categorization and the arithmetic average in regression. The deliberate randomization of the cut-point and attributes paired with ensemble averaging should be able to minimize variance more strongly than other approaches' weaker randomization schemes [47]. In addition, using the whole original learning sample rather than bootstrap clones reduces bias. In terms of computational complexity, assuming balanced trees, the process is on the order of $N \log N$ in terms of learning sample size. The Extra Trees method, like Random Forest, is unaffected by the value utilized, despite being a critical hyperparameter to control. The parameters used for building Extra Trees classification model are $N_{\text{estimators}} = 100$, $\text{Criteria} = \text{entropy}$, $\text{Max features} = 2$.

4.4.6 Ensemble Model

Machine learning ensemble methods utilize the insights gained from different learning models to help make more accurate and better judgments. For example, noise, variation, and bias are the most common error causes in learning K -Nearest Neighbors, random forests, stochastic gradient descent classifier and an extra tree classification model [38]. As a result, the suggested ensemble technique achieves

considerable accuracy, outperforming all individual classifiers [48]. Furthermore, it is an improvement approach used to the outputs of several algorithms to improve model's accuracy. Ensemble approaches in machine learning assist in reducing these error-causing elements, ensuring that machine learning (ML) methods are accurate and stable. This research employed the hybridization of Decision Trees, resulting in a final classification that is superior to individual classifiers. As seen in Fig. 8, the graphic displays the recommended ensemble model.

4.4.7 Prediction Assessment Parameters

The confusion matrix sometimes referred to as the error matrix, is represented by a matrix detailing how a classification model performed on a test data set. The confusion matrix represents counts between the expected and actual values as shown in Fig. 9. True negative (TN) indicates how many false positives were removed from the training set. Similarly, the number of correctly identified positive cases is denoted by the abbreviation "TP," which stands for True Positive. False Positive indicates the number of genuine negative cases incorrectly labeled positive. In contrast, False Negative indicates the opposite: the number of genuine positive examples that were incorrectly labeled negative [49, 50].

After creating the confusion matrix, the efficacy of the data classification methods was compared using the metrics accuracy rate, recall, F1 measure, and RMSE (root mean squared error). The following parameters are determined:

- Accuracy: The metric accuracy is used to quantify the efficacy of a classifier. The number of properly classified values in a set is termed as accuracy, and it is determined using an equation [51].

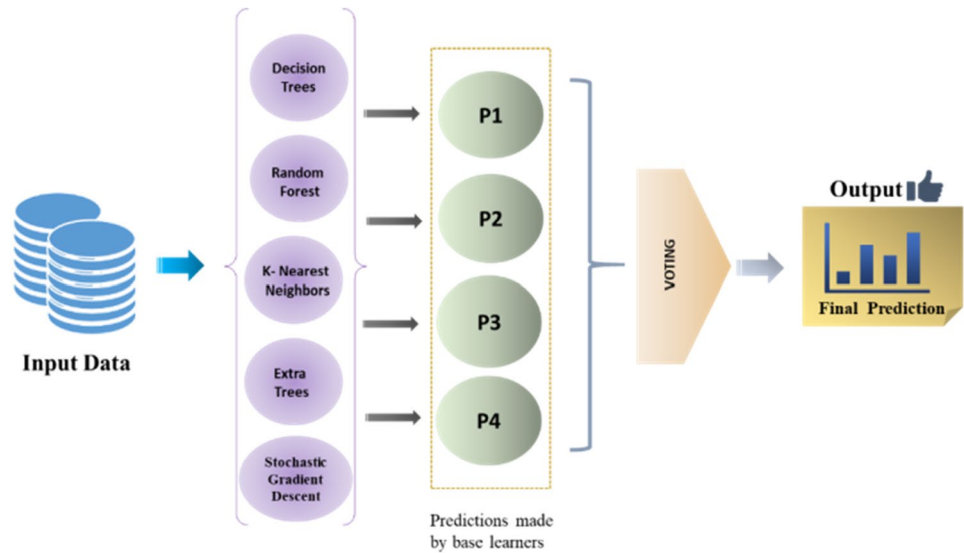
$$\text{Accuracy} = \frac{TP_n + TN_n}{TP_n + TN_n + FP_n + FN_n} \quad (7)$$

- Root Mean Squared Error (RMSE): RMSE is considered as the difference between the values predicted by the model and the values actually observed. N denotes the number of observations. The formula for RMSE is given in Eq. (3).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N ((\text{Predicted}_i - \text{Actual}_i)^2)}{N}} \quad (8)$$

- Recall: A statistic that measures how many patients were accurately identified as having a disease relative to the overall number of patients with the disease. The perception of recall is the number of patients diagnosed with the disease. Sensitivity is another term for recall. TP stands for "truly positive".

Fig. 8 Proposed ensemble model



		Actual values	
		Positive	Negative
Predicted values	Positive	True-Positive	False-Positive
	Negative	False-Negative	True-Negative

Fig. 9 Confusion matrix

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- Precision: Precision, also known as a positive predictive value, refers to the proportion of really positive outcomes relative to the total number of such outcomes that may be anticipated. Thus, precision may be defined as the rate at which positive values are accurately identified: precision is expressed mathematically as shown in Eq. (5).

$$P_n = \frac{TP}{TP + FP} \tag{10}$$

- F1-Score (F1): The F1 score, often known as the F-score or F-measure [52, 53], is the harmonic-mean of accuracy and sensitivity as given in Eq. (6).

$$F1 = \frac{2 \times Precision \times Sensitivity}{(Precision + Sensitivity)} \tag{11}$$

5 Simulation Results

The assessment must be organized and provide visible, clear findings that may be utilized and improved. Data analysis and evaluation are critical parts of the evaluation process, and there are a variety of assessment methodologies to choose from. Several criteria were used to assess the effectiveness of classification techniques. The tools and libraries needed for model training are shown in Table 4.

5.1 Confusion Matrix Results

A confusion matrix may be used to assess a classifier's potential. Correctly categorized outcomes are represented by all diagonal elements, whereas off diagonals represent misclassified outcomes. Therefore, a confusion matrix with just diagonal entries and all other elements set to zero will be the best classifier. After the categorization procedure, a confusion matrix yields actual and expected values. Table 5 depicts the confusion matrices of several classification methods.

5.2 Results of Machine Learning Classifiers

The accuracy of the six machine learning classifiers for food-borne disease prediction on the same dataset without using any scaling strategies is shown in Table 6. The Ensemble Learning approach outperformed the other five techniques, except for *Salmonella* and *Listeria* diseases, when we employed an unscaled dataset. The ensemble method offers the highest accuracy, with a 97.74% detection rate for Norovirus, 99.64% detection rate for *Clostridium*, 99.0%

detection rate for *Campylobacter* and 98.99% detection rate for *E. coli*. On the other hand, stochastic gradient descent (SGD) fared severely in virtually all instances. The K-nearest neighbor classifier outperformed the competition with 99.63% accuracy in predicting Salmonella sickness, whereas the Extremely Randomized Tree method came in second with 99.61%. In Listeria disease, Extremely Randomized Tree had the highest accuracy of 98.73%, while Stochastic Gradient Descent had the lowest accuracy of 86.86.

The performance of various metrics is illustrated in Figs. 10 and 11a–d without using any scaling technique, including root mean squared error, accuracy, recall and F1-score. Applying multiple measurements, ensemble learning and decision trees were able to obtain excellent outcomes in all metrics; on the other hand, stochastic gradient descent achieved the lowest score in all illnesses.

5.3 MinMax Scaling and Standard Scaling Methods

In this investigation, we utilized two scaling techniques (minmax and standard); however, the minmax and standard scaling methods improved the overall performance of the classifiers. This is because particular models (though not all) are scale-sensitive, meaning that they give disproportionate weight to characteristics that appear on bigger scales. Furthermore, scaling provides a good framework since it aligns all features, increasing the likelihood that the model will recognize the proper patterns. As depicted in Tables 7, 8, 9, 10, 11 and Figs. 12, 13, 14, 15, the outcomes of accuracy, RMSE and precision were obtained for various classifiers after applying the minmax and standard scalars. As shown in previous Table 6, when we compare the results of different classifiers using minmax and standard scaling strategies with the results of classifiers that did not use a scaling strategy, we observe that the ensemble approach employing scaling techniques is more accurate in nearly predicting all food borne pathogens.

The root mean squared error (RMSE) error measure is used extensively and is an effective all-around error measure for numerical predictions. This metric informs us how accurate our estimations are and how much they deviate from

the actual data. The RMSE score of our proposed ensemble learning (EL) strategy is the best, indicating that EL correctly predicted the data, but the Stochastic Gradient Descent model fared poorly. Ensemble model excelled by achieving predicted accuracies of 98.84 percent in norovirus illness 99.75 percent in campylobacter disease and 99.83 percent in listeria disease using the standard Scaler approach. Using minmax scaler, the ensemble model had outstanding results, with projected accuracy rate of 99.94 percent for salmonella disease, 99.64 percent for Clostridium disease and 99.99 percent for *E. coli* disease. The stochastic gradient descent model did not perform well on both Scaling techniques, with the lowest accuracy (58%), recall (67%) and F1 Score (72%). Other tree-based techniques, such as decision trees and random forest, yield high prediction scores, suggesting strong performance. The decision tree classifier achieved the second most excellent prediction results. The results indicate that stochastic gradient descent (SGD) did not yield satisfactory RMSE scores in either Standard or Minimax scaling methods. However, our proposed method, ensemble learning (EL), demonstrated consistent and accurate data predictions regarding RMSE scores. Notably, illnesses including Norovirus, *Salmonella*, *Clostridium*, *Campylobacter*, *E. coli* and *Listeria* have RMSE scores ranging from 0.120 to 0.320 when using normal and MinMax scaling as shown in Tables 8 and 9.

This study presents a new approach to ensemble learning techniques that has not been previously explored in the existing literature. This study proposes using ensemble learning to enhance the precision and dependability of detecting and forecasting foodborne illnesses, including their symptoms and diagnostics. The research presented in this study showcases the efficacy of ensemble learning through empirical evidence of its superior performance compared to other established methodologies. As seen in Table 12, proposed ensemble learning method obtained more than 90% accuracy detecting in all food-borne disease pathogens. This study provides a quantitative evaluation of the impact of ensemble learning on disease prediction metrics, specifically RMSE scores.

6 Discussion

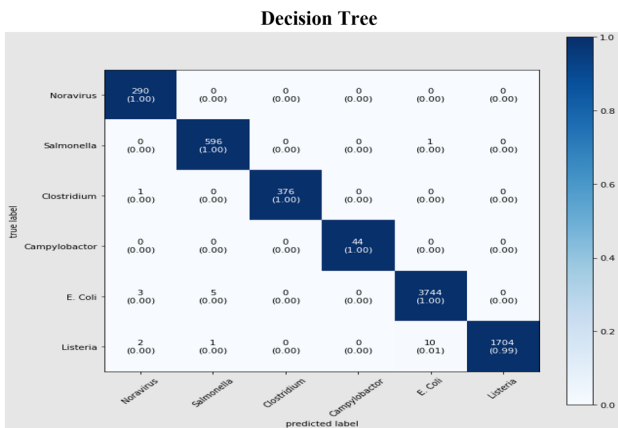
One of the most significant phenomena anticipated to impact food safety in the upcoming years is the presence (or emergence) of unexpected or new pathogens in foods. Food pathogen identification is essential for reasons of public food safety and health. Contagious pathogens found in contaminated food are distributed widely in the environment. People who work in the food industry are exposed to various food pathogens that may cause fatal food-borne illnesses such as hemolysin dysfunction, fever, diarrhea

Table 4 Model-training prerequisites

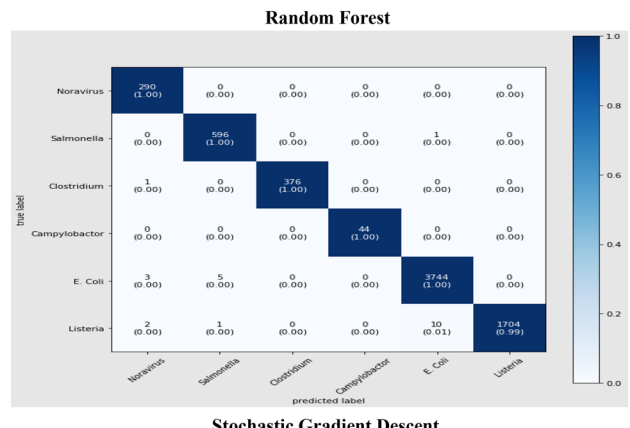
Implementation requirements	
Processor	Intel i7(recommended)
Ram	32 GB
GPU	4 GB
Windows	10
Software	https://www.anaconda.com/
Web-based interactive environment	Jupyter notebook
Programming language	Python

Table 5 Confusion matrix of classification models

Decision tree

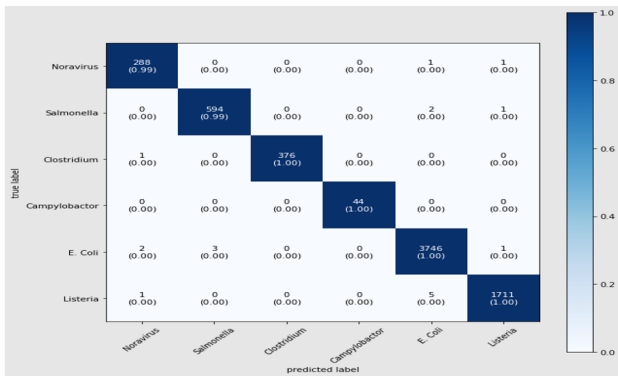


Random forest



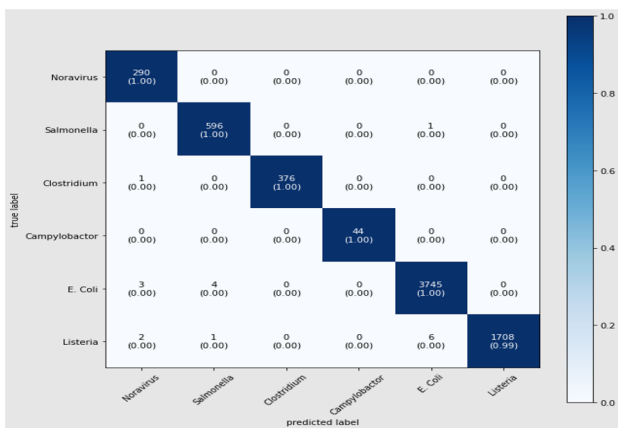
k-Nearest Neighbor

k-nearest neighbor

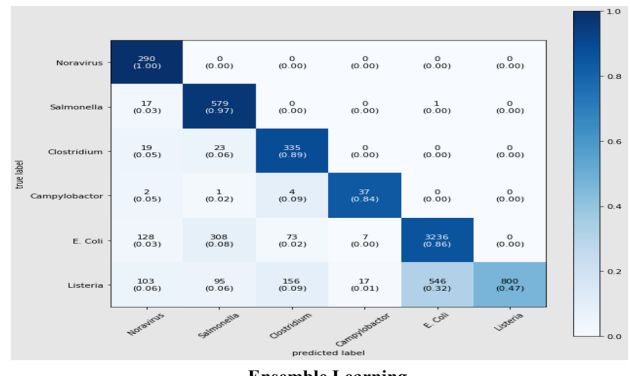


Extremely Randomised Tree

Extremely randomised tree

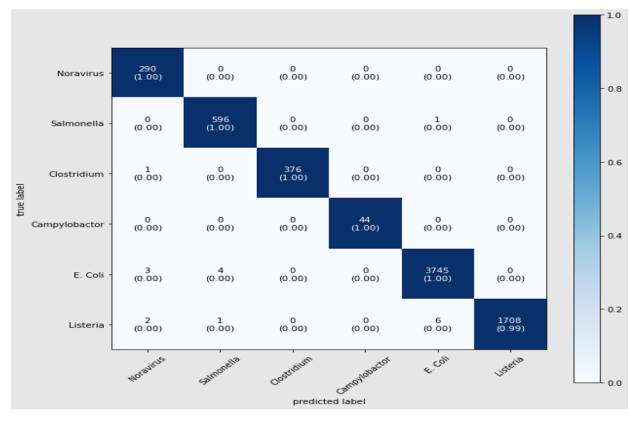


Stochastic gradient descent



Ensemble Learning

Ensemble learning



and stomach cramps. It is crucial to monitor outbreaks of food-borne illness to spot patterns in the foods, regions and pathogens involved. In this field, it is necessary to know the genotype and subtype of food contamination strains to

identify the transmission source, define, and compare variants. Additionally, various strains of food-borne pathogens are connected with human illness differently. These variations may be ascribed, among other things, to the hardness

Table 6 Performance evaluations based on accuracy without any scaling technique

Diseases	Norovirus	Salmonella	Clostridium	Campylobacter	E. coli	Listeria
Algorithms						
Decision tree (DT)	97.41	95.41	92.51	98.61	98.41	96.61
Random forest (RF)	95.66	99.46	97.61	97.13	99.56	96.44
K-nearest neighbor (KNN)	97.73	99.63	98.99	98.73	97.53	96.73
Stochastic gradient descent (SGD)	87.00	81.66	87.16	75.66	75.83	86.86
Extremely randomised tree	96.44	99.61	96.43	97.73	98.00	98.73
Ensemble learning (EL)	97.74	92.41	99.64	99.00	98.99	93.94

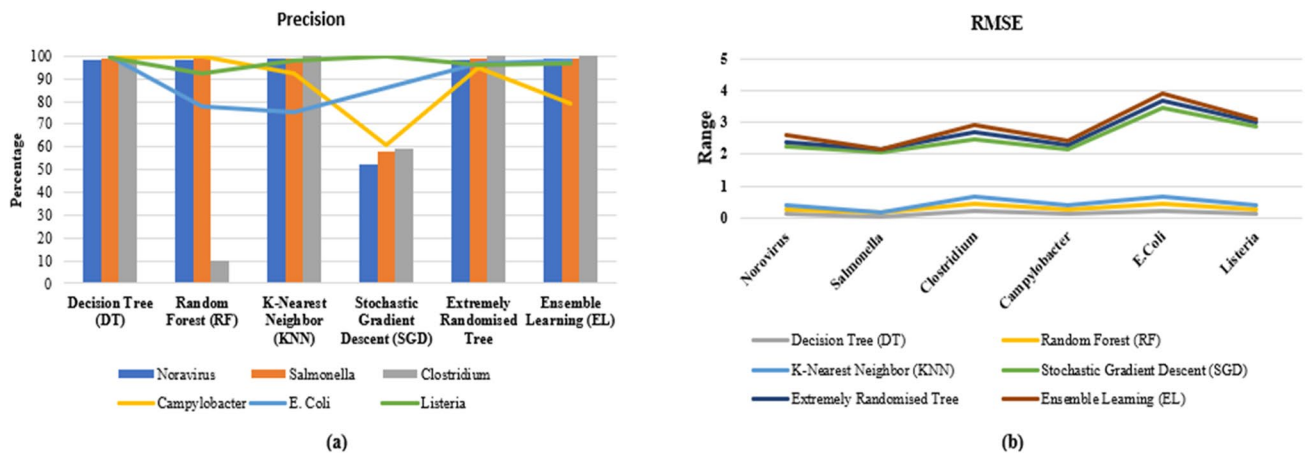


Fig. 10 Results of precision and RMSE (without any scaler)

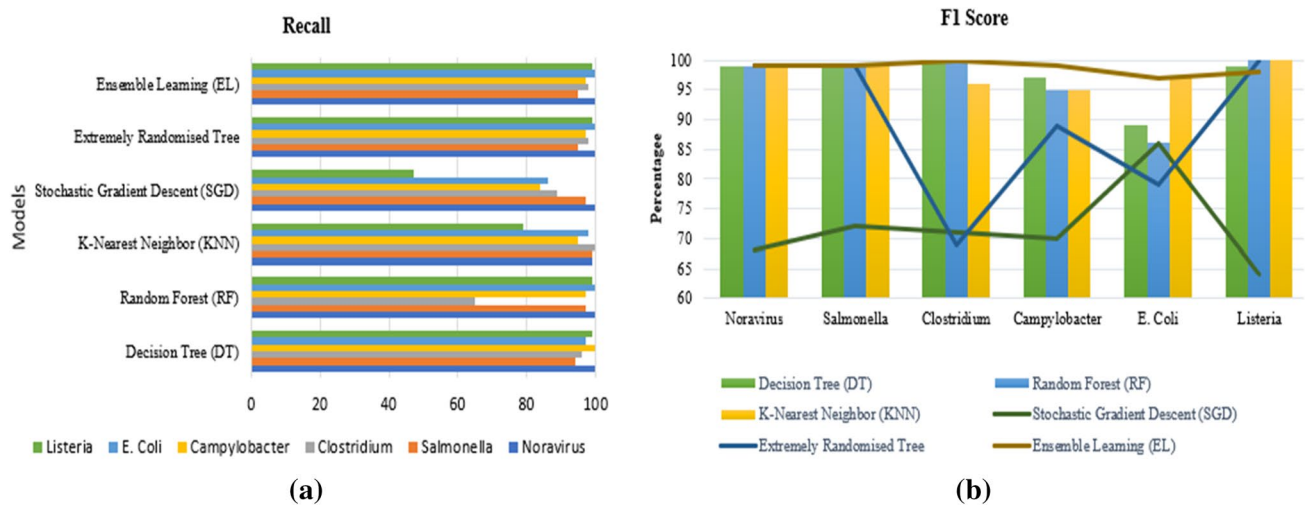


Fig. 11 Results of recall and F1 score (without any scaler)

Table 7 Models accuracy-based scaling methods (MinMax and standard scaler)

Diseases→ Algorithms↓	Norovirus			Salmonella			Clostridium			Campylobacter			E. coli			Listeria		
	MinMax scaler	Standard scaler	MinMax scaler	MinMax scaler	Standard scaler	MinMax scaler	Standard scaler	MinMax scaler	Standard scaler	MinMax scaler	Standard scaler	MinMax scaler	Standard scaler	MinMax scaler	Standard scaler	MinMax scaler	Standard scaler	
Decision tree (DT)	98.61	97.61	99.81	96.81	92.51	97.51	92.51	97.60	98.61	98.61	97.60	99.51	96.51	96.61	98.51	96.61	98.51	
Random forest (RF)	94.66	96.66	99.46	98.46	97.66	97.66	98.66	97.67	98.66	98.66	97.67	99.56	98.56	96.66	99.56	96.66	99.56	
K-nearest neighbor (KNN)	97.73	98.73	99.63	99.63	98.93	98.93	97.93	97.76	98.73	98.73	97.76	99.53	97.53	96.73	94.63	96.73	94.63	
Stochastic gradient descent (SGD)	87.86	86.86	77.76	87.76	85.16	85.16	86.16	70.80	79.86	79.86	70.80	77.96	79.96	89.86	86.76	89.86	86.76	
Extremely randomised tree	98.73	98.73	99.43	98.43	96.43	96.43	97.43	98.78	97.73	97.73	98.78	99.83	98.83	98.73	99.83	98.73	99.83	
Ensemble learning (EL)	98.74	98.84	99.94	90.64	99.64	99.64	98.84	99.75	98.74	98.74	99.75	99.99	97.24	93.74	92.94	93.74	92.94	

Table 8 Metrics RMSE and precision-based MinMax scaling methods

Diseases→ Algorithms↓	Norovirus		Salmonella		Clostridium		Campylobacter		E. coli		Listeria	
	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision
Decision tree (DT)	0.133	98	0.023	99	0.233	100	0.133	100	0.233	100	0.133	99
Random forest (RF)	0.126	98	0.025	99	0.226	100	0.126	100	0.226	100	0.126	99
K-nearest neighbor (KNN)	0.123	99	0.020	99	0.223	100	0.123	100	0.223	100	0.123	100
Stochastic gradient descent (SGD)	1.879	52	1.579	58	1.779	59	1.779	61	1.779	86	1.479	100
Extremely randomised tree	0.124	98	0.014	99	0.224	100	0.124	100	0.224	100	0.124	100
Ensemble learning (EL)	0.220	99	0.010	99	0.220	100	0.120	100	0.220	100	0.120	100

Table 9 Metrics F1 and recall-based MinMax scaling methods

F1 and recall (MinMax scaler)												
Diseases→ Algorithms↓	Norovirus		<i>Salmonella</i>		<i>Clostridium</i>		<i>Campylobacter</i>		<i>E.Coli</i>		<i>Listeria</i>	
	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall
Decision tree (DT)	99	100	99	100	100	100	100	100	100	100	99	99
Random forest (RF)	99	100	99	100	100	100	100	100	100	100	99	99
K-nearest neighbor (KNN)	99	98	99	99	100	100	100	100	100	100	100	100
Stochastic gradient descent (SGD)	68	100	72	97	71	89	70	84	86	86	69	47
Extremely randomised tree	99	100	99	100	100	100	100	100	100	100	100	99
Ensemble learning (EL)	99	100	99	100	100	100	100	100	100	100	100	99

of specific strains that allows them to live and multiply in food-related situations or to their higher virulence towards people. In this study, we used a food-borne disease dataset to predict various pathogens in food-borne diseases. Data preprocessing and exploratory data analysis were used to clean the data and categorize illnesses by state, hospital, and fatality. To aggregate various classification models to predict food-borne illnesses, we used machine learning models and suggested an ensemble learning model built via majority voting. Predictions of pathogens are evaluated using quality metrics, including precision, accuracy, recall, F-score, root mean square error and confusion matrix. In order to diagnose food-borne illness infections, we analyzed data on cases and pathogens. In addition, we used machine learning to assess connections between geographical location, period, and potentially contaminated foods. We examined the findings of multiple models to find the most accurate pathogen prediction model.

7 Conclusion

Emerging food-borne microorganisms, such as bacteria, viruses and parasites, are undoubtedly one of the most significant food safety concerns affecting the food industry and public health agencies. Emerging pathogen research focuses on improving techniques for identifying and managing evolving infections, shortening the period between a pathogen's appearance and its management, and anticipating new food safety issues. The problem of food-borne pathogens' emergence is expected to be effectively controlled mainly via the establishment and implementation of robust and effective surveillance systems. These

programs will allow early detection and study of developing (or reemerging) food-borne infectious diseases and deploy effective control and preventative strategies. Lastly, it is anticipated that the creation and use of innovative molecular tools for researching food-borne diseases would assist in a better understanding of the aspects that led to the rise of these infections. In recent years, there has been an increase in the prevalence of diseases transmitted by ingesting contaminated food. A more effective strategy for reducing the spread of foodborne infections is detecting potential pathogenic bacteria in food and processing settings. Analytical results and machine learning-based pathogen classification may aid in identifying and treating foodborne illnesses. This study employed machine learning models to aggregate various classification models to predict food-borne infections and a suggested ensemble learning model was developed using majority voting. Using quality criteria such as accuracy, precision, recall, F-score, root mean square error, and confusion matrix, the applied models, predict and assess food-borne illnesses. The ensemble model achieves the best prediction average accuracy, 99.74 percent, followed by the extra randomized tree classifier. According to the RMSE, precision, recall and F1 score as shown in various Figures and Tables, the suggested ensemble techniques outperformed KNN and other ensemble approaches. This study concludes that machine learning-based approach can be utilized for predicting the pathogens in food-borne illnesses. Thus, these models, especially ensemble-based approaches can be endorsed as benchmark models for prediction modeling. We may enhance our work in the future in two ways. First, we may enhance the average effectiveness of predictions for food-borne pathogen detection by using several feature

Table 10 Metrics RMSE and precision-based standard scaling methods

Diseases→ Algorithms↓	RMSE and precision (standard scaler)											
	Norovirus		Salmonella		Clostridium		Campylobacter		E. coli		Listeria	
	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision
Decision tree (DT)	0.233	97	0.023	99	0.213	100	0.130	94	0.223	100	0.333	100
Random forest (RF)	0.226	98	0.025	99	0.216	98	0.129	100	0.216	98	0.326	100
K-nearest neighbor (KNN)	0.223	99	0.020	99	0.213	100	0.129	98	0.213	100	0.323	97
Stochastic gradient descent (SGD)	1.779	62	1.579	58	1.719	69	1.777	61	1.759	86	1.379	98
Extremely randomised tree	0.224	98	0.014	99	0.214	100	0.123	97	0.214	100	0.324	94
Ensemble learning (EL)	0.120	99	0.010	99	0.210	100	0.129	96	0.210	97	0.320	95

Table 11 Metrics F1 and recall-based standard scaling methods

Diseases→ Algorithms↓	RMSE and precision (standard scaler)											
	Norovirus		Salmonella		Clostridium		Campylobacter		E.Coli		Listeria	
	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision	RMSE	Precision
Decision tree (DT)	0.233	98	0.023	99	0.213	100	0.130	100	0.223	100	0.333	100
Random forest (RF)	0.226	98	0.025	99	0.216	100	0.129	100	0.216	100	0.326	100
K-nearest neighbor (KNN)	0.223	99	0.020	100	0.213	100	0.129	100	0.213	100	0.323	100
Stochastic gradient descent (SGD)	1.779	52	1.579	58	1.719	59	1.777	61	1.759	86	1.379	100
Extremely randomised tree	0.224	98	0.014	100	0.214	100	0.123	100	0.214	100	0.324	100
Ensemble learning (EL)	0.120	99	0.010	99	0.210	99	0.129	100	0.210	100	0.320	100

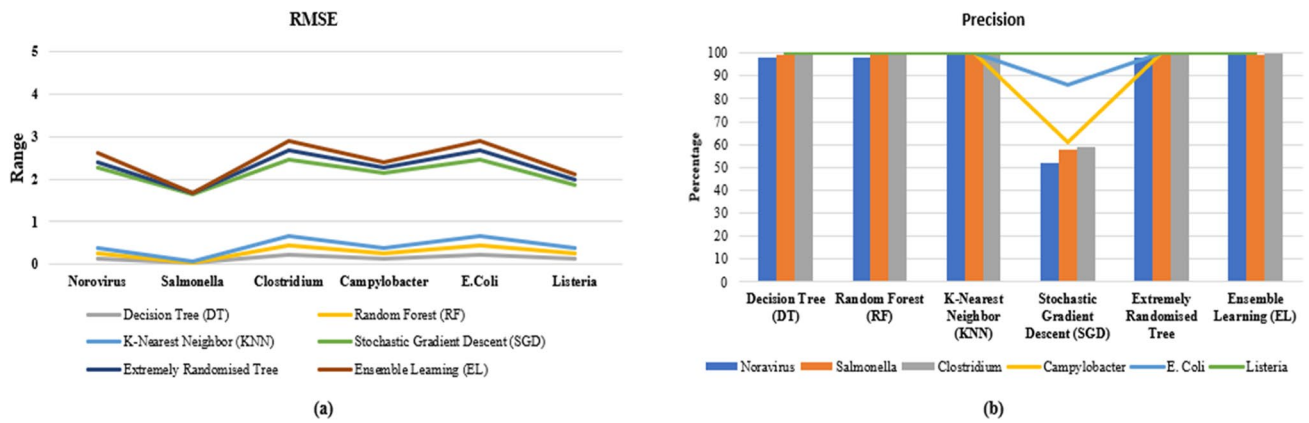


Fig. 12 Results of RMSE and precision (with MinMax scaler)

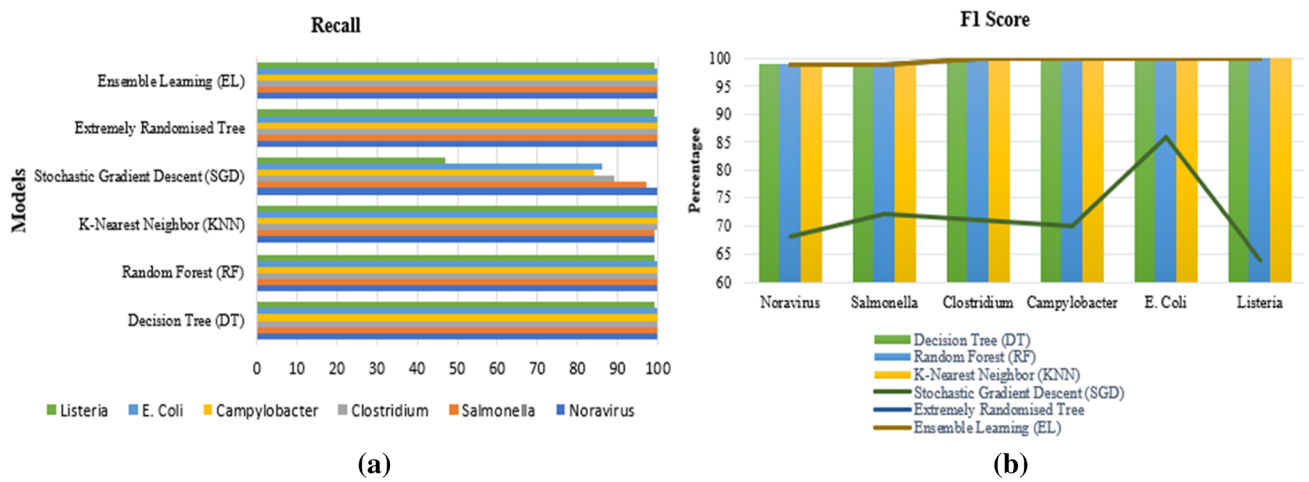


Fig. 13 Results of recall and F1 score (with MinMax scaler)

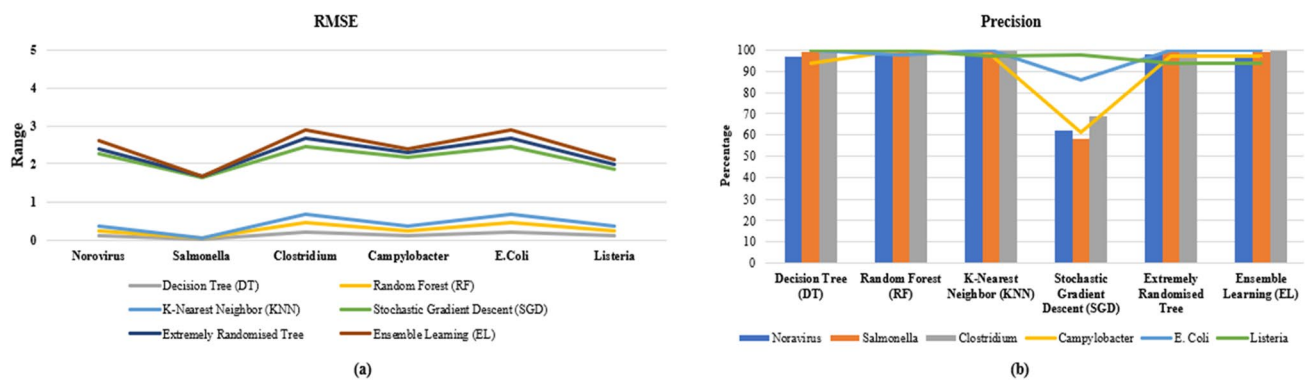


Fig. 14 Results of RMSE and precision score (with standard scaler)

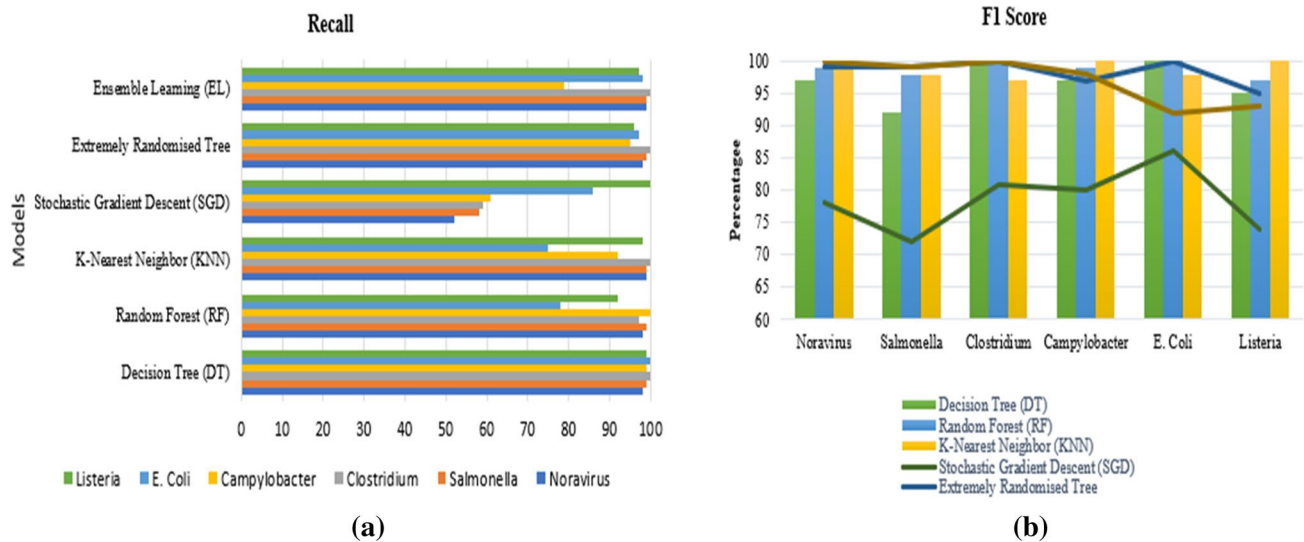


Fig. 15 Results of recall and F1 score (with standard scaler)

Table 12 Comparison with existing work

References	Datasets	Techniques used	Findings
[24]	Data taken from agriculture pond	kNN, ANN and SVM	k-NN outperformed with the accuracy of 77.15%
[9]	NFDSRS	RF, DT, GBDT, adaptive boosting techniques	GBDT model achieved the accuracy of 69%
[26]	https://figshare.com/s/7a3ededa8cedd95b9fb7	SVM	SVM obtained accuracy of 83% in the case of <i>Escherichia coli</i>
[28]	NGS data	RF, SVM, GBM, NN and LB	SVM outperformed with 89% accuracy
[33]	Previously conducted research investigation	Boosted gradient classifier	Boosted gradient classifier scored a median accuracy score of 64%
Proposed approach (Ensemble learning)	Kaggle	Decision trees (DT), random forests (RF), k-nearest neighbors (k-NN), stochastic gradient descent (SGD) and extremely randomized trees	Accuracy obtained more than 90% in all food-borne pathogens

selection strategies and optimization procedures. Second, we may extend our model to incorporate more food-borne pathogens, which would assist in diagnosing a variety of diseases.

Data Availability Not applicable.

Declarations

Competing interests Non-financial interests and authors have no conflict of interests.

References

1. Finger JA, Baroni WS, Maffei DF, Bastos DH, Pinto UM (2019) Overview of foodborne disease outbreaks in Brazil from 2000 to 2018. *Foods* 8(10):434
2. Sharif MK, Javed K, Nasir A (2018) Foodborne illness: threats and control. Foodborne diseases. Academic Press, Cambridge, pp 501–523
3. Jung Y, Jang H, Matthews KR (2014) Effect of the food production chain from farm practices to vegetable processing on outbreak incidence. *Microb Biotechnol* 7(6):517–527
4. Kaur I, Garg R, Kaur T, Mathur G (2023) Using artificial intelligence to predict clinical requirements in healthcare. *J Pharm Negat Results* 2023:4177–4180
5. Vidyadharani G, Vijaya Bhavadharani HK, Sathishnath P, Ramanathan S, Sariga P, Sandhya A, Sugumar S (2022) Present and pioneer methods of early detection of food borne pathogens. *J Food Sci Technol* 59(6):2087–2107

6. Torgerson PR, Devleesschauwer B, Praet N, Speybroeck N, Willingham AL, Kasuga F, de Silva N (2015) World Health Organization estimates of the global and regional disease burden of 11 foodborne parasitic diseases, 2010: a data synthesis. *PLoS Med* 12(12):e1001920
7. Vilne B, Meistere I, Grantiņa-Ieviņa L, Ķibilds J (2019) Machine learning approaches for epidemiological investigations of food-borne disease outbreaks. *Front Microbiol* 10:1722
8. Kadariya J, Smith TC, Thapaliya D (2014) *Staphylococcus aureus* and staphylococcal food-borne disease: an ongoing challenge in public health. *BioMed Res Int* 2014:1–9
9. Wang H, Cui W, Guo Y, Du Y, Zhou Y (2021) Machine learning prediction of foodborne disease pathogens: algorithm development and validation study. *JMIR Med Inform* 9(1):e24924
10. Pandey SK, Bhandari AK (2023) A systematic review of modern approaches in healthcare systems for lung cancer detection and classification. *Archiv Comput Methods Eng* 30:1–20
11. Kumar Y, Gupta S (2023) Deep transfer learning approaches to predict glaucoma, cataract, choroidal neovascularization, diabetic macular edema, drusen and healthy eyes: an experimental review. *Archiv Comput Methods Eng* 30(1):521–541
12. Heredia N, García S (2018) Animals as sources of food-borne pathogens: a review. *Animal nutrition* 4(3):250–255
13. Saravanan A, Kumar PS, Hemavathy RV, Jeevanantham S, Kamalesh R, Sneha S, Yaashikaa PR (2021) Methods of detection of food-borne pathogens: a review. *Environ Chem Lett* 19:189–207
14. Chukwu EE, Nwaokorie FO, Coker AO, Avila-Campos MJ, Ogunisola FT (2019) 16S rRNA gene sequencing: a practical approach to confirming the identity of food borne bacteria. *IFE J Sci* 21(3):13–25
15. Koul A, Bawa RK, Kumar Y (2023) Artificial intelligence techniques to predict the airway disorders illness: a systematic review. *Archiv Comput Methods Eng* 30(2):831–864
16. Hu W, Feng K, Jiang A, Xiu Z, Lao Y, Li Y, Long Y (2020) An in situ-synthesized gene chip for the detection of food-borne pathogens on fresh-cut cantaloupe and lettuce. *Front Microbiol* 10:3089
17. Nesakumar N, Lakshmanakumar M, Srinivasan S, Jayalatha Jbb A, Balaguru Rayappan JB (2021) Principles and recent advances in biosensors for pathogens detection. *ChemistrySelect* 6(37):10063–10091
18. Zheng S, Yang Q, Yang H, Zhang Y, Guo W, Zhang W (2023) An ultrasensitive and specific ratiometric electrochemical biosensor based on SRC-A-CRISPR/Cas12a system for detection of *Salmonella* in food. *Food Control* 146:109528
19. Chenar SS, Deng Z (2021) Hybrid modeling and prediction of oyster norovirus outbreaks. *J Water Health* 19(2):254–266
20. Zhang P, Cui W, Wang H, Du Y, Zhou Y (2021) High-efficiency machine learning method for identifying foodborne disease outbreaks and confounding factors. *Foodborne Pathog Dis* 18(8):590–598
21. Chenar SS, Deng Z (2018) Development of artificial intelligence approach to forecasting oyster norovirus outbreaks along Gulf of Mexico coast. *Environ Int* 111:212–223
22. Min HJ, Mina HA, Deering AJ, Bae E (2021) Development of a smartphone-based lateral-flow imaging system using machine-learning classifiers for detection of *Salmonella* spp. *J Microbiol Methods* 188:106288
23. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, Davis JJ (2018) Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal *Salmonella*. *bioRxiv* 2018:380782
24. Polat H, Topalcengiz Z, Danyluk MD (2020) Prediction of *Salmonella* presence and absence in agricultural surface waters by artificial intelligence approaches. *J Food Saf* 40(1):e12733
25. Amado TM, Bunuan MR, Chicote RF, Espenida SMC, Masangcay HL, Ventura CH, Enriquez LAC (2019) Development of predictive models using machine learning algorithms for food adulterants bacteria detection. 2019 IEEE 11th international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management (HNICEM). IEEE, New York, pp 1–6
26. Lupolova N, Dallman TJ, Holden NJ, Gally DL (2017) Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb Genom*. <https://doi.org/10.1099/mgen.0.000135>
27. Hiura S, Koseki S, Koyama K (2021) Prediction of population behavior of *Listeria monocytogenes* in food using machine learning and a microbial growth and survival database. *Sci Rep* 11(1):1–11
28. Njage PMK, Henri C, Leekitcharoenphon P, Mistou MY, Hendriksen RS, Hald T (2019) Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. *Risk Anal* 39(6):1397–1413
29. Borujeni MS, Ghaderi-Zefrehei M, Ghanegolmohammadi F, Ansari-Mahyari S (2018) A novel LSSVM based algorithm to increase accuracy of bacterial growth modeling. *Iran J Biotech* 16(2):105
30. Bandoy DJ, Weimer BC (2020) Biological machine learning combined with campylobacter population genomics reveals virulence gene allelic variants cause disease. *Microorganisms* 8(4):549
31. Hill AA, Crotta M, Wall B, Good L, O'Brien SJ, Guitian J (2017) Towards an integrated food safety surveillance system: a simulation study to explore the potential of combining genomic and epidemiological metadata. *Royal Soc Open Sci* 4(3):160721
32. Maharana A, Cai K, Hellerstein J, Hswen Y, Munsell M, Staneva V, Nsoesie EO (2019) Detecting reports of unsafe foods in consumer product reviews. *JAMIA Open* 2(3):330–338
33. Olm MR, Bhattacharya N, Crits-Christoph A, Firek BA, Baker R, Song YS, Banfield JF (2019) Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. *Sci Adv* 5(12):eaax5727
34. Nogales A, Morón RD, García-Tejedor AJ (2020) Food safety risk prediction with Deep Learning models using categorical embeddings on European Union data. Preprint at <https://arxiv.org/abs/2009.06704>
35. Ahsan MM, Mahmud MA, Saha PK, Gupta KD, Siddique Z (2021) Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* 9(3):52
36. Rudra T, Paul P (2021) Heart disease prediction using traditional machine learning.
37. Kaur I, Sandhu AK, Kumar Y (2022) A hybrid deep transfer learning approach for the detection of vector-borne diseases. 2022 5th international conference on contemporary computing and informatics (IC3I). IEEE, New York, pp 2189–2194
38. Peng T, Chen X, Wan M, Jin L, Wang X, Du X, Yang X (2021) The prediction of hepatitis E through ensemble learning. *Int J Environ Res Public Health* 18(1):159
39. Nogales A, Díaz-Morón R, García-Tejedor AJ (2022) A comparison of neural and non-neural machine learning models for food safety risk prediction with European Union RASFF data. *Food Control* 134:108697
40. Wheeler NE (2019) Tracing outbreaks with machine learning. *Nat Rev Microbiol* 17(5):269–269
41. Martínez-García PM, López-Solanilla E, Ramos C, Rodríguez-Palenzuela P (2016) Prediction of bacterial associations with plants using a supervised machine-learning approach. *Environ Microbiol* 18(12):4847–4861
42. Bhardwaj P, Bhandari G, Kumar Y, Gupta S (2022) An investigational approach for the prediction of gastric cancer using artificial

- intelligence techniques: a systematic review. *Archiv Comput Methods Eng* 29:1–22
43. Lumogdang CFD, Wata MG, Loyola SJS, Angelia RE, Angelia HLP (2019) Supervised machine learning approach for pork meat freshness identification. *Proceedings of the 2019 6th international conference on bioinformatics research and applications*. ACM, New York, pp 1–6
 44. Chowdhury NH, Reaz MBI, Haque F, Ahmad S, Ali SHM, Bakar AAA, Bhuiyan MAS (2021) Performance analysis of conventional machine learning algorithms for identification of chronic kidney disease in type 1 diabetes mellitus patients. *Diagnostics* 11(12):2267
 45. Kader MS, Ahmed F, Akter J (2021) Machine learning techniques to precaution of emerging disease in the poultry industry. *2021 24th international conference on computer and information technology (ICCIT)*. IEEE, New York, pp 1–6
 46. Rani P, Kumar R, Jain A (2021) Coronary artery disease diagnosis using extra tree-support vector machine: ET-SVMRBF. *Int J Comput Appl Technol* 66(2):209–218
 47. Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, Bukhari SAC (2019) An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access* 7:54007–54014
 48. Weller DL, Love T, Wiedmann M (2021) Interpretability versus accuracy: a comparison of machine learning models built using different algorithms, performance measures, and features to predict *E. coli* levels in agricultural water. *Front Artif Intell* 4:19
 49. Goyal P, Gopala Krishna DN, Jain D, Rathi M (2021) Foodborne disease outbreak prediction using deep learning. *Innovations in computational intelligence and computer vision*. Springer, Singapore, pp 165–172
 50. Kaur I, Kumar Y, Sandhu AK, Ijaz MF (2023) Predictive modeling of epidemic diseases based on vector-borne diseases using artificial intelligence techniques. *Computational intelligence in medical decision making and diagnosis*. CRC Press, Boca Raton, pp 81–100
 51. Singh PD, Kaur R, Singh KD, Dhiman G (2021) A novel ensemble-based classifier for detecting the COVID-19 disease for infected patients. *Inf Syst Front* 23(6):1385–1401
 52. Kaur I, Sandhu AK, Kumar Y (2022) Artificial intelligence techniques for predictive modeling of vector-borne diseases and its pathogens: a systematic review. *Archiv Comput Methods Eng* 29:1–31
 53. Kaur I, Sandhu AK, Kumar Y (2021) Analyzing and minimizing the effects of Vector-borne diseases using machine and deep learning techniques: a systematic review. *2021 sixth international conference on image information processing (ICIIP)*, vol 6. IEEE, New York, pp 69–74

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.