IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **Volume 11, Iss 4 , 2022**

A Review on Various Food Diseases by Using Machine Learning Algorithm

Dr. Savya Sachi

Assistant Professor, Department of Information Technology, L. N. Mishra College of Business Management Muzaffarpur, Bihar

Abstract- Public health is still threatened by food safety. Large, newly available data sets may be used by machine learning to enhance food supply safety and lessen the effects of food safety events. genomes of food borne pathogens and new data streams, such as transactional, text, and trading data, have observed new uses made possible by machine learning strategy, like the forecasting of antibiotic resistance, the attribution of sources to pathogens, risk assessment, and the identification of foodborne outbreaks. Within this In this post, we give a thorough introduction to machine learning with a focus on food safety, along with a summary of current advancements and uses. Despite the fact that many of these applications are still in their infancy, general and domain-specific machine learning pitfalls and challenges are starting to be identified and addressed. These developments are crucial for the potential use and future deployment of large data sets and the machine learning models that go along with them for food safety applications.

Keywords— *Public Health, Machine Learning Algorithm, Transactional, Text, Trading Data, Food Diseases.*

INTRODUCTION

The phrase "machine learning" was first used by Arthur Samuel (1959, p. 211) to describe how a computer could learn to play and win at checkers in a fashion that "if done by human beings or animals, would be described as involving the process of learning." Later, Samuel's concept was expanded to include the field of study that develops artificial intelligence (AI) in computers without requiring explicit programming. Its application dates back to the early 1700s, when sailors needed assistance navigating the ocean, astronomers and geodesists developed least-squares methods to describe planetary orbits based on measurements (data) (Stigler 1986). The theory and tools of modern machine learning were blueprinted by visionaries like Alan Turing after World War II (Turing 1950). Some of the most widely used algorithms and models, such as nearest neighbors, random forests, and neural networks, were invented between the 1960s and the 1990s. After the term Big Data was popularized in both the scientific community and the general public around the 1990s and 2000s, the explosive growth of machine learning benefited from vastly increasing data sizes, exponentially growing computer power, and new refinements of old tools, eventually leading to a myriad of breakthroughs. Notable milestones include recognition of handwritten digits (LeCun et al. 1989) and speech (Hochreiter & Schmidhuber 1997); classification of objects such as cats, dogs, and planes (Krizhevsky et al. 2012); and mastery of gameplay without human knowledge in the game of Go (Silver et al. 2017). As a subfield of artificial intelligence, machine learning differs from traditional algorithmic problem-solving by not attempting to program an exhaustive list of explicit instructions or rules. Instead, a machine learning system learns from examples and generalizes to new cases based on their closeness to learned examples (instance-based learning) or trains a model with data to learn its parameters through optimization and makes predictions using new (test) data (model-based learning). The data-driven and rule-agnostic characteristics of machine learning make it attractive for certain types of tasks. First, for problems that are difficult to pose mathematically (Shardanand & Maes 1995) or without explicit solution algorithms, machine learning may find a reasonable approximation. Second, some problems are so complex for existing methods that a prohibitively long list of rules would need to be programmed for their solutions. For example, for the game of Go, it is combinatorically unrealistic to find the optimal move through a brute-force search. Third, some tasks must cope with new data to which hard-coded rules are impossible to adapt, such as detecting novel

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

spam in emails and on social media. Finally, machine learning can provide insights and verify heuristics on large-scale problems, such as the Go strategies and tactics innovated and rediscovered by DeepMind's AlphaGo (Baker & Fan 2017)

LITERATURE REVIEW

Foodborne infections continue to pose a serious and persistent threat to public health. According to Scallan et al. (2011), foodborne illness affects 48 million Americans, or 1 in 6 of the population, and results in 128,000 hospital admissions as well as 3,000 fatalities annually. Food safety was identified as a focus area in the 2020 vision released by the US Healthy People initiative in 2010 (Koh 2010). As of 2019, the Foodborne Diseases Active Surveillance Network (FoodNet) surveillance data showed that none of the vision's goals for controlling six key foodborne pathogens by 2020 had been achieved. Over more than a century, major transformations in food production, distribution, and regula tion have taken place, driven by and feeding into macrosocietal trends such as population increase, urbanization, and globalization (Doyle et al. 2015, Phillips 2006). Massive changes and advances in the food industry and supply chains have generated large volumes of data, especially in recent years, similar to in other sectors and industries. A plethora of data has been explored in inno vative ways and at different stages along the farm-to-table continuum to improve the safety of the food supply. For instance, at preharvest, terrain and meteorological data were investigated for predicting pathogen contamination on produce farms (Strawn et al. 2013), and in the retail set ting, paperless auditing and record keeping enabled 1.4 million monthly measurements of internal cooking temperatures of rotisserie chickens for food safety assurance (Yiannas 2015). At the end of the food supply chain, consumer interactions with foods, including transaction, consumption, and experience feedback and sharing, also create copious amounts of data. These novel data streams (NDS) are increasingly propagated and accessible via digital platforms such as social media, search histories, crowdsourcing sites, and consumer reviews and commentary, as well as databases of product sales and consumption records. Mining of these data to inform food safety and public health is on the horizon (Harris et al. 2014, Maharana et al. 2019). On the surveillance front, data-intensive systems play important roles in tracking foodborne illness cases and agents. Examples at the US federal level include PulseNet (Swaminathan et al. 2001), the National Antimicrobial Resistance Monitoring System (NARMS) (Gupta et al. 2004, Zhao et al. 2006), FoodNet (Scallan & Mahon 2012), and the National Outbreak Reporting Sys tem (Hall et al. 2013). Data collected by some of these systems have surged in the recent decade owing to the incorporation of genomic data on foodborne pathogens. Implementation of whole genome sequencing (WGS) in surveillance and outbreak investigation has fueled an explosion of publicly available foodborne pathogen genomes in new systems such as GenomeTrakr (Allard al Center for Biotechnology Informa tion's Pathogen Detection (https://www.ncbi.nlm.nih.gov/pathogens). Routine use of WGS in public health microbiology has given rise to a data-driven area known as genomic epidemiology (Deng et al. 2016). Recent advances in the data science approach to food safety have led to the discussion of Big Data (Marvin et al. 2017), a term that is not traditionally associated with food safety. To meet analytical challenges created by the deluge of data, machine learning has emerged as a promising tool for data-intensive analytics in food safety. In April 2019, the Food and Drug Administration (FDA) released a statement on "steps to usher the US into a new era of smarter food safety," in which artificial intelligence and machine learning applications in food safety were proposed (Sharpless & Yiannas 2019). Given the rapid emergence of machine learning applications in food safety, we aim to pro vide a comprehensive overview of the new field by introducing fundamentals of the methodology, reviewing recent and notable progress, and discussing challenges and potential pitfalls. Machine learning, as a general-purpose data analytics tool, has been used in other areas of agricultural and food science, such as food processing and quality evaluation, as reviewed elsewhere (Du & Sun 2006). In this review, we focus on domain-specific applications in food safety and public health.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

MACHINE LEARNING METHODS

During training, machine learning systems may receive guidance or supervision. Based on the amounts of supervision provided, the learning can be categorized into supervised, unsupervised, semi-supervised, and reinforced.

(i) In typical supervised learning tasks, such as classification and regression, the training data fed to the learning system are labeled with the desired outcome or the ground truth. To build a cat/dog classifier, a training set of many pet images must be assembled and labeled with the classes: cats, dogs, or neither. To develop a regressor that predicts a continuous numeric value, such as housing prices given a set of features (e.g., neighborhood, size, year built), many instances of houses are collected to fit a regression model, each including both features and a label: its price.

(ii) In unsupervised learning, training data are unlabeled, leaving the algorithm to unearth the hid den patterns. An example is the identification of customer groups through behavioral/transactional data without an a priori defined grouping. Another important application is anomaly and novelty detection. For example, a learning system shown mostly normal network traffic can learn to detect cyber intrusions.

(iii) Labeling of data can be labor intensive and is not readily available for large data sets. There are often few labeled instances among many unlabeled examples. As a combination of supervised and unsupervised learning, semi-supervised algorithms can weigh in on unlabeled data's contribution to feature-target relations, usually taking advantage of the assumption that nearby samples are likely to share the same labels (Zhu et al. 2003). For example, in automatic speech recognition, accented speech is commonly underrepresented in training data and problematic for supervised learning. Semi-supervised learning of tone and pitch accent has been shown to reduce the need for labeled training data for speech recognition (Levow 2006).

(iv) Unlike supervised learning, in which training data comes with specific answers to the question (class labels), reinforcement learning relies on a learning system (agent) to find the best strategy or path (policy) in a given situation. The learning is achieved through a trial-and-error process during which the agent is rewarded or penalized by the actions it takes, with the goal of maximizing the reward over time. Reinforcement finds plenty of use in robotics and gaming, from robots learning to walk (Haarnoja et al. 2018) to the AlphaGo program beating the world Go champion (Silver et al. 2017).

EXAMPLES OF ALGORITHMS

Numerous machine learning algorithms have been developed that vary in sophistication to accom modate problems of different levels of complexity. Four representative and fundamental learning algorithms are summarized in Figure 1. K-means is an unsupervised algorithm that partitions similar observations into clusters dynamically. It uses geometric centers of observations (centroids) to prototype clusters, and an observation is then assigned to a cluster if it is closer to the cluster's centroid than any other centroid (Figure 1a). A support-vector machine (SVM) uses planes to best separate observations of different classes by representing them as points. A new observation is mapped onto the space, and its class is pre dicted according to the side of the plane on which it falls. SVM is particularly efficient for data in which different classes are well separated (Figure 1b). Decision trees (Figure 1c) attempt to split instances into different classes recursively through the interaction of different features. A new sample follows particular branches determined by the features to land on a leaf that provides its predicted class. It is possible to randomize this approach through averaging the results from a multitude of different trees (random forest) and to grow trees by following a more quantifiable criterion through the introduction of an objective function (gradient boosting). Both methods are examples of ensemble learning. An artificial neural network (ANN) simulates biological neural networks by comprising layers of interconnected artificial neurons called processing units

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 11, Iss 4 , 2022

(Figure 1d). These units or nodes receive input information and process it through a system that includes a linear combination of weights and input, a nonlinear activation function, and output signals to the next layer of nodes. Each ANN consists of one input layer, one or more intermediate layers called hidden layers, and one output layer. Together, they convert initial inputs into results for regression or classification tasks. An ANN containing many hidden layers is called a deep neural network, which is the core of deep learning.

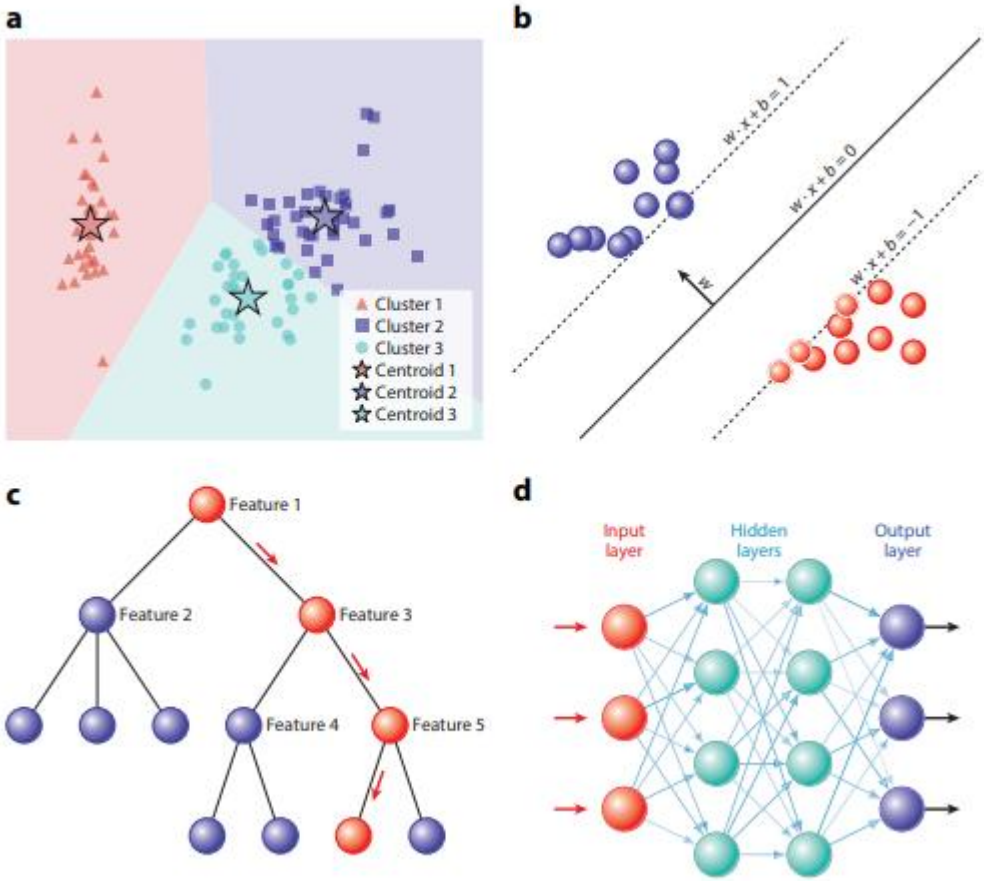

Figure 1- Examples of machine learning models. (a) A decision boundary plot of k-means clustering with three clusters, with new samples being grouped to a cluster by the colored region it lands on. (b) A line dividing two classes in a support-vector machine with a certain margin. w contains the trainable parameters, and x stands for the vector representation of a sample. (c) A decision tree with five features. A sample is classified into a certain class following the red arrow. (d) A neural network with two hidden layers; the arrow stands for a connection between units, with transparency indicating the connection strength

MACHINE LEARNING APPLICATIONS USING GENOMIC DATA

In genetics, genomics, and medicine, machine learning holds promise for making biological dis coveries and predictions from large genomic data sets.Machine learning models have been trained to recognize patterns and elements in DNA sequences, a process known as sequence annotation (Libbrecht & Noble 2015). Genomic signatures or biomarkers have been identified via machine learning techniques to assist in disease diagnosis, clinical decision-making, and drug discovery and development (He et al. 2019, Vamathevan et al. 2019). One could assume that machine learning analysis of foodborne pathogen genomes is an iteration or extension of established methodology and therefore straightforward owing to the relatively small genomes of the pathogens. However, domain-specific opportunities and challenges continue to arise as machine learning is increasingly used to tap into the rapidly growing resources of foodborne pathogen

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

genomes and their meta data. Still in their infancy, such applications have been focused on antimicrobial resistance (AMR) prediction and genomic source attribution of certain pathogens.

(i) Antimicrobial Resistance Prediction- Measurement of antimicrobial resistance or susceptibility traditionally relies on phenotypic assays that measure growth inhibition of an antibacterial agent on a population of pure culture bacte ria. A common technique for antimicrobial susceptibility testing (AST) is broth dilution, which involves a range of antibiotic concentrations and determines the minimum inhibitory concentra tion (MIC) of a drug to inactivate or inhibit the growth of a particular bacterial isolate. Clinical breakpoints are assigned to divide AST results into categories in correlation with the likelihood of treatment outcome, including susceptible (high probability of a favorable outcome), resistant (low probability of a favorable outcome), and sometimes intermediate (Humphries et al. 2019, Turnidge & Paterson 2007).

(ii) Genomic Source Attribution of Foodborne Pathogens- According to the Centers for Disease Control and Prevention, approximately 95% of foodborne illnesses in the United States are sporadic, non-outbreak cases whose food exposures and con tamination sources are challenging to determine. With source information for most foodborne infections being largely unknown, it is difficult to understand foodborne illness epidemiology and develop intervention measures to prevent and mitigate such illnesses.Major foodborne pathogens, such as Salmonella and E. coli, are zoonotic enteric bacteria whose primary reservoirs include live stock and wild animals. Unlike AMR, for which many genetic determinants have been identified and characterized, mechanistic understanding of zoonotic host specificity and tropism is still lim ited. The lack of genetic markers prevents a rule-based approach to source prediction but creates an opportunity for machine learning investigations that do not necessarily require a priori knowl edge of genetic determinants of bacterial host preference or adaptation.

(iii) Challenges and Potential Pitfalls of Machine Learning Applications in Food Safety Using Genomic Data- AST and source attribution represent related but different types of phenotypic inference from genomic data. When machine learning inference is intended to identify genetic variations causally associated with specific phenotypes, it can be considered as a subset of microbial genome-wide association studies (mGWAS). Adapted from GWAS methods used in human genetics, mGWAS face challenges and pitfalls specific to bacterial species, including genome-wide linkage disequi librium and strong population structuring, such as distinct lineages and clonal groups (Eyre et al. 2017, San et al. 2019). Such genetic and population traits can lead to identification of genotype– phenotype associations that are correlational but not causal. The still-nascent use of machine learning in food safety genomics has only begun to consider such challenges and pitfalls, either during feature selection (Lupolova et al. 2019) or at results confirmation (Drouin et al. 2016).

CONCLUSION

Although there aren't many exciting uses of machine learning with NDS in the field of food safety now, there are many that could be modeled after related fields. As Kaufman et al. (2014) pioneered using retail sales data, loyalty cards (Aiello et al. 2019), restaurant sales, online grocery (Huyghe et al. 2017) or delivery (Schulz et al. 2019) data sets, which have been used in consumer behavior and nutrition applications, could be applied to identify likely outbreak food vehicle sources. Additional data, such as product-specific characteristics (e.g., shelf life, probable consumption date, the likelihood that a specific product contains a particular pathogen), or variables influencing the purchase of specific items (e.g., weather, holidays, sporting events), features frequently used in retail consumption demand forecasting models, could be added to the prediction task. Markets and restaurants are examples of places where contaminated products may have been purchased. Aggregated credit card transactions, which have been used to construct machine learning models of consumer shopping trajectories (Krumme et al. 2013, Singh et al. 2015), could be utilized to identify locations. There are numerous uses for location data from smartphone applications or cell phone call data records in researching the transmission of infectious diseases from person to person (Oliver et al. 2020). However, there are currently few examples of foodborne disease applications (Sadilek et al.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **Volume 11, Iss 4 , 2022**

2017, Teyhouee et al. 2017). Strategies utilizing social media and search queries could be extended beyond the surveillance of foodborne illnesses to encompass other aspects of food safety such as product recalls, allergies, or food safety laws.

REFERENCES

[1]. Dodd C, Aldsworth T, Stein R, Cliver D, Riemann H. In: Jones JL, editor. Foodborne Diseases. Netherlands: Academic Press; 2017.
[2]. Bean H, Griffin P, Goulding JS. Foodborne disease outbreaks, 5-year summary, 1983-1987. Journal of Food Protection 1990;53(8):711-728. [doi: 10.4315/0362-028x-53.8.711]
[3]. Oliver SP. Foodborne pathogens and disease special issue on the national and international PulseNet network. Foodborne Pathog Dis 2019 Jul;16(7):439-440. [doi: 10.1089/fpd.2019.29012.int] [Medline: 31259613]
[4]. Liu J, Bai L, Li W, Han H, Fu P, Ma X, et al. Trends of foodborne diseases in China: lessons from laboratory-based surveillance since 2011. Front Med 2018 Feb 27;12(1):48-57. [doi: 10.1007/s11684-017-0608-6] [Medline: 29282610]
[5]. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med 2015 Dec 3;12(12):e1001921 [FREE Full text] [doi: 10.1371/journal.pmed.1001921] [Medline: 26633831]
[6]. Centers for Disease Control and Prevention. Foodborne Diseases Active Surveillance Network, 1996. MMWR Morb Mortal Wkly Rep 1997 Mar 28;46(12):258-261 [FREE Full text] [Medline: 9087688]
[7]. Foodborne Disease Monitoring and Reporting System. National Center for Food Safety Risk Assessment. URL: https:/ /foodnet.cfsa.net.cn/ [accessed 2021-01-16]
[8]. Oldroyd RA, Morris MA, Birkin M. Identifying methods for monitoring foodborne illness: review of existing public health surveillance techniques. JMIR Public Health Surveill 2018 Jun 06;4(2):e57 [FREE Full text] [doi: 10.2196/publichealth.8218] [Medline: 29875090]
[9]. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M, Roy SL, et al. Foodborne illness acquired in the United States major pathogens. Emerg Infect Dis 2011 Jan;17(1):7-15. [doi: 10.3201/eid1701.p11101]
[10]. Mandal P, Biswas A, Choi K, Pal U. Methods for rapid detection of foodborne pathogens: an overview. American Journal of Food Technology 2011 Jan 15;6(2):87-102. [doi: 10.3923/ajft.2011.87.102]
[11]. Law JW, Ab Mutalib N, Chan K, Lee L. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front Microbiol 2014 Jan 12;5:770 [FREE Full text] [doi: 10.3389/fmicb.2014.00770] [Medline: 25628612]
[12]. Naravaneni R, Jamil K. Rapid detection of food-borne pathogens by using molecular techniques. J Med Microbiol 2005 Jan;54(Pt 1):51-54. [doi: 10.1099/jmm.0.45687-0] [Medline: 15591255]
[13]. Pan W, Zhao J, Chen Q. Classification of foodborne pathogens using near infrared (NIR) laser scatter imaging system with multivariate calibration. Sci Rep 2015 Apr 10;5:9524 [FREE Full text] [doi: 10.1038/srep09524] [Medline: 25860918]
[14]. Flint JA, Van Duynhoven YT, Angulo FJ, DeLong SM, Braun P, Kirk M, et al. Estimating the burden of acute gastroenteritis, foodborne disease, and pathogens commonly transmitted by food: an international review. Clin Infect Dis 2005 Sep 01;41(5):698-704. [doi: 10.1086/432064] [Medline: 16080093]
[15]. Kuehn BM. Agencies use social media to track foodborne illness. JAMA 2014 Jul 09;312(2):117-118. [doi: 10.1001/jama.2014.7731] [Medline: 24963655]
[16]. Sadilek A, Kautz H, DiPrete L, Labus B, Portman E, Teitel J, et al. Deploying Nemesis: preventing foodborne illness by data mining social media. AIMag 2017 Mar 31;38(1):37-48. [doi: 10.1609/aimag.v38i1.2711]
[17]. Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, et al. Discovering foodborne illness in online restaurant reviews. J Am Med Inform Assoc 2018 Dec 01;25(12):1586-1592 [FREE Full text] [doi: 10.1093/jamia/ocx093] [Medline: 29329402]