5-2023

Detecting Pathobiomes Using Machine Learning

Valerie Jackson
University of Arkansas, Fayetteville

Valerie Jackson

DETECTING PATHOBIOMES USING MACHINE LEARNING

DETECTING PATHOBIOMES USING MACHINE LEARNING

A thesis submitted in partial fulfillment
of the requirements for the degree of
Honors Bachelor of Science in Industrial Engineering

By

Valerie Jackson
University of Arkansas
Bachelor of Science in Industrial Engineering, 2023

April 2023
University of Arkansas

This thesis is approved for recommendation to the Graduate Council

---

Thesis Director: Dr. Chase Rainwater

Committee member: Dr. John L. Kent

Abstract

Machine learning is a field with high growth potential due to the overall continuous progressions, developments, advancements, and improvements caused by the way it is used to help interpret and use large amounts of data [1]. One type of data that can be collected and analyzed by these machine learning models is data that is associated with DNA and information that the DNA gives. The research will be focusing specifically on using machine learning technology to detect pathobiomes indicative of salmonella pork. The pathobiome associated with salmonella is very similar to others, and this causes a problem for classification/detection with short-read sequencing platforms [2]. Because of this, it is important for decision makers to understand what kind of data is needed to help accurately predict these pathobiomes. This research consists of a variety of experiments that help determine what this kind of data is. This is done by reading data taken from various sequences from The National Center for Biotechnology Information(NCBI) database. This project is also being conducted within the backdrop of an existing project from the Walmart foundation project, Improving Food Safety of Pork Supply Chain [3].

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Problem Introduction

China is the world's number one producer of pork, exceeding the production level of the next-leading country four times over [4]. China also leads in pork consumption, exceeding other countries levels by a similar amount [5]. The more pork produced and consumed, the more room there is for growth and spread of food-borne diseases. There are over 200 different types of these diseases caused by eating contaminated food. These diseases could include bacterial infections, parasites, viral infections, chemical infections, and more. Within the span of one year, these diseases occur in nearly 1/10 of the world's population [6]. Around 0.07 percent of food-borne illnesses end in death, and these deaths affect young children disproportionately, meaning that 3/10 deaths occur in children under 5 years of age [6]. In China specifically, pathogenic bacterial outbreaks cause 70 percent of food-borne diseases. In addition, salmonella ranks first out of these pathogenic bacteria in China [7].

Detecting salmonella before reaching the consumer is a priority for maintaining food safety throughout China and the rest of the world. Existing detection methods typically rely on sampling pork product at different stages of the supply chain and completing salmonella prevalence testing in a laboratory. While this type of testing is reliable, it is slow to complete and requires significant financial resources in lost product and laboratory technicians. Moreover, the time spent to complete sampling for salmonella in the laboratory tests prohibits timely safety mitigation and tracing. This work seeks to develop an artificial intelligence-based detection model that makes use of RNA sequences that can be obtained in numerous ways without destroying food product [8]. Our efforts specifically seek to identify the pathobiome, a set of organisms associated with reduced health status in a certain host, in each sample of DNA from the salmonella [9]. The pathobiome is important because it plays a key role in the signs and symptoms of disease that

are observed in the infected person [10]. Pathobiomes are a more accurate and specific way of identifying disease than by simply referring to the sickness as the outcome of the effects of a single virus or pathogen [10]. In order to benefit from pathobiomes, we seek to develop a machine learning-based model to classify and identify salmonella, and determine how different factors of the inputted data affect the model's ability to predict salmonella. These results of the samples and pathobiome information will have the ability to inform managers in charge of risk to determine what practical food safety interventions need to be implemented [11].

The broader objective of this research is to understand how machine learning can be used to detect the pathobiome of a strand of DNA. In the process of doing this however, we illustrate the fact that an artificial intelligence model can be integrated as a food safety risk tool. We show that a machine learning model can monitor alongside biosensors more traditional detection approaches to improve the confidence in contamination assessment and enable closer to real-time mitigation strategies. The specific contributions of specific research is the improved understanding the role of RNA sequence length, number of RNA samples and number of pathogens considered impacts the accuracy of machine learning classification in our problem's context.

2 Research Methodology

To begin on our preliminary research for this topic, we used the seqviewer tool from MATLAB. This seqviewer tool is a part of what is called the "Bioinformatics Toolbox" [12]. This is an extension of MATLAB that reads genomic and proteomic data from standard file formats such as SAM, FASTA, CEL, and CDF, as well as from online databases such as the NCBI Gene Expression Omnibus and GenBank [13]. This tool allows the user to input a given accession number, and get the full nucleotide sequence of the data shown in Figure 2.1.

We can see the "Annotated CDS" as well to see the protein coding part of the nucleotide sequence. This tool also gives the ability to use the "find word" feature to search a specific nucleotide pattern (for example "AGT") and it reveals everywhere in the sequence that occurs. We can see the ORFs (open reading frames) of the sequence as well. An ORF is "a span of genomic 'letters' that falls between the start and stop signals. Researchers can scan the genome for open reading frames to find genes that encode proteins."[14]. The user can also see the complement and reverse complement of the sequence as well. The "comments" tab gives additional information on the article where the data is cited, authors, as well as what organism it was found in and what type of nucleotide (For example: "Salmonella enterica subsp. enterica serovar Bovismorbificans strain").

The next step was to identify situations where DNA strands have been used as inputs for deep learning models before. One way is by creating the DNA stand into images to visualize the strands and can then help us to turn them into predictions for what pathobiomes those strands represent [15]. This approach takes DNA sequence data, trains amodel to fit it, and outputs images that help to predict the sequence. This idea was more on track with what we wanted to implement, however, was difficult for us to show prediction for a specific pathobiome. Therefore, we proceeded to explore a TensorFlow based solution for the problem. An example

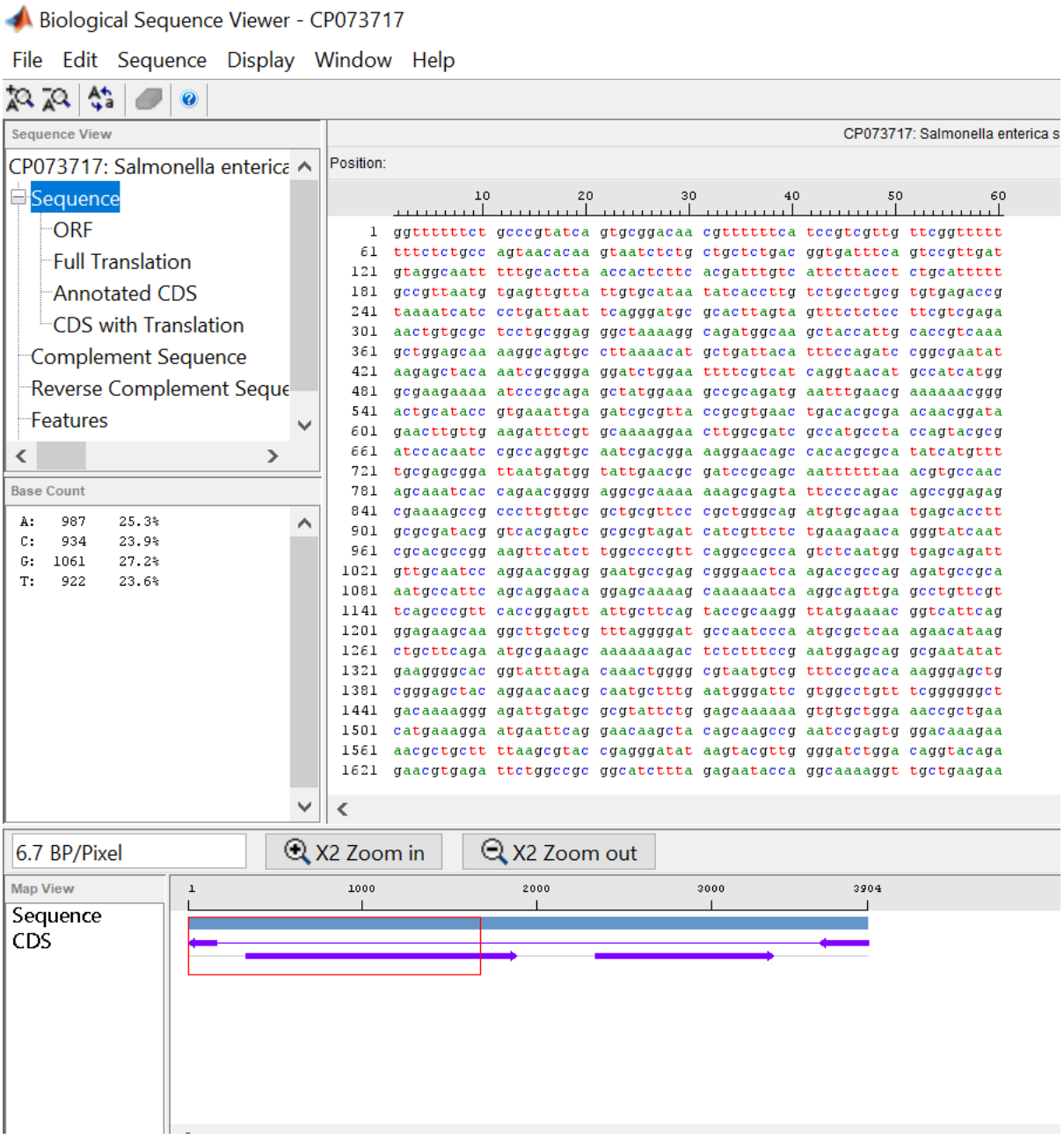

Figure 2.1: Sequence from Bioinformatics Toolbox

```
def DNA_matrix(seq):
    tem2 = ['[aA]', '[cC]', '[gG]', '[tT]']
    for i in range(len(tem2)):
        ind = [m.start() for m in re.finditer(tem2[i], seq)]
        tem = np.zeros(len(seq), dtype=int)
        tem[ind] = 1
        if i == 0:
            a = np.zeros((len(seq), 4))
        a[..., i] = tem
    return a
```

Figure 2.2: Code for One-Hot Matrix

```
[[0. 1. 0. 0.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]
 [0. 0. 1. 0.]
 [0. 0. 0. 1.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [0. 1. 0. 0.]
```

Figure 2.3: Outputted Matrix

developed to identify whether a sequence of DNA given is human or non-human was previously publised [16]. This example reads in two different data files to train the code. The first one being that for the sequence itself containing 10,000 human sequences and 10,000 random sequences generated by RSAT random-seq [16]. The next file was for the labels and it was simply a series of 1s and 0s where 1 stands for human sequences and 0 stands for random sequences.

The implemementation began by reading in the data, and then converting it into one-hot matrix shown in Figure 2.2 and 2.3. Then the machine learning model is trained with layers. The first layer is a one dimensional convolution. This type

of layer "...accepts a multichannel one dimensional signal, convolves it with each of its multichannel kernels, and stacks the results together into a new multichannel signal that it passes on to the next layer" [17]. The next layer is a second one dimensional convolution. The third layer is a maximum pooling layer. This layer is a pooling operation that calculates the maximum value in each batch of the data [18]. The next layer is a flatten layer that converts the data into a 1-dimensional array for inputting it to the next layer. [19]. Then a dense layer is added "...that is deeply connected with its preceding layer which means the neurons of the layer are connected to every neuron of its preceding layer" [20]. The last layer to be added is a dropout layer. How this works is "During training, some number of layer outputs are randomly ignored or 'dropped out.' This has the effect of making the layer look-like and be treated-like a layer with a different number of nodes and connectivity to the prior layer [21]. Next, the model is eventually evaluated. During the evaluation, a testing and loss is computed for the user to view. We can then retrieve and examine the weight matrix, and convert it into a PWM matrix [16]. After running the code using a standard laptop, and fully understanding how this process worked, we moved forward to begin on our adaptation of this method with the goal of identifying specific pathobiomes for salmonella.

To begin with our own methodology, we needed to find a source to use that we knew was taken from a sample of salmonella and that we knew was in pork to train our data model. As we previously stated in our abstract, this project is being conducted within the backdrop of an existing project from the Walmart foundation project, Improving Food Safety of Pork Supply Chain [3]. However, access to data that we would need from this project (like salmonella DNA sequences) was not available to us. Due to this, we decided to use the sequences of salmonella that were taken from a strain isolated from dried pork sausage associated with an outbreak in France [22]. We decided on this source due to the fact that it was from a food product that was also pork.

We utilized the seqviewer tool from the Bioinformatics Toolbox that we described earlier to extract the sequence used in this source. The sequence came

out as a list of letters that represent the four bases of a DNA molecule. These being adenine (A), guanine (G), thymine (T), and cytosine (C). These bases are what bond together in a DNA molecule to give its structure [23]. Our data was a long combination of "A"s, "G"s, "T"s, and "C"s that we copied into a new file that we would use in our version of the python code described in the section above.

We wanted to first verify that we could get a similar model to the one that identified human and non-human DNA sequences up and running with just the salmonella data from the pork sausage example. To do this, we made the label file, in this case, a series of 66 "1"s that are listed down to represent every salmonella sequence (there are 66 salmonella sequences in this source). Unlike in the human and non-human example, We only had one label identifier in this case since we simply wanted to see if the code worked in the way we hoped. After some tweaking, we were able to get the code to run, and it seemed like it was time to push the model beyond what has already been done.

We wanted to see if we could extend this model to be able to identify whether a sequence of DNA belonged to a certain pathobiome of a food-borne disease. To begin this process, we did more research into food-borne diseases as a whole. According to the Center for Disease Control and Prevention (CDC), the top 5 germs causing deaths from food eaten in the United States are Salmonella, toxoplasma gondii, listeria monocytogenes, norovirus, and campylobacter [24]. With this in mind, we began to look for literature regarding outbreaks of each of these diseases. The motivation for this was with hopes that we could find accession numbers for DNA sequences of these diseases that we could input into the seqviewer to produce the full DNA sequence usable to us.

The sequences that we decided to use for toxoplasma gondii were taken from feces of domestic cats in Colombia [25]. This source contained 37 sequences. The sequences of Listeria monocytogenes were taken from a source that dealt with the rational design of DNA sequence-based strategies for subtyping listeria monocytogenes [26]. This source contained 30 sequences. The sequences of norovirus taken from patients [27]. This source had 126 sequences. Finally, the sequences

of campylobacter were taken from a source dealing with the characterization of gyrA mutations associated with fluoroquinolone resistance in campylobacter coli by DNA sequence analysis and MAMA PCR [28]. This source contained 53 sequences. For each of these sources, we inputted the accession number into the seqviewer extension to output the full DNA sequence.

Each of these sequences was added to a single text file named "Research Final Combined Sequence" which combined all of the sequences for salmonella, toxoplasma gondii, listeria monocytogenes, norovirus, and campylobacter together. We then created a text file to make labels for this combined file. We named the file "Research Label" and listed out a series of numbers to represent each disease type. Due to the fact that some of the sequences from these sources were shorter than the full length, we cut a couple of each type out for the data to fit into the model we would create. The file had 304 numbers total, 65 "1"s representing salmonella, 36 "2"s representing toxoplasma gondii, 25 "3"s representing listeria monocytogenes, 126 "4"s representing norovirus, and 52 "5"s representing campylobacter.

Using a similar process to that outlined in the human vs non-human example [16], we trained the model using the files we created, "Research Final Combined Sequence" and "Research Label". We did the same thing that was done previously, where the code began by reading in the data given and then converting it into one-hot matrix. Then, the model is trained and eventually evaluated. During the evaluation, testing and loss are reported for the user to view [16].

3 Experiments and Analysis

Once we knew that this extended version ran without any issues, we began to conduct an analysis of what impact the length of the strands, the number of outcomes, and the number of rows of each strand had on the resulting testing loss and accuracy values of the prediction. The length of the strands simply means how many bases (letters 'a', 'g', 't', or 'c') are included in each strand. The number of outcomes refers to how many different choices the resulting prediction could be. In the case of two outcomes, this is referring to the original example, with the outcomes being either human or non-human [16]. In the case of five outcomes, this is where we are using the data from the five different pathobiomes that we mentioned earlier. Those are salmonella, toxoplasma gondii, listeria monocytogenes, norovirus, and campylobacter. For the case of three outcomes, we simply used the same data for the first three pathobiomes (salmonella, toxoplasma gondii, and listeria monocytogenes). The number of rows is referring to how many different strands of the given length are in the data set. The two metrics that we used to determine the model's ability to predict the outcome were testing loss and accuracy. Testing loss is defined as "...the penalty for a bad prediction. That is, loss is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater." [29]. It is also important to note that loss is subjective to the data type that we are using. Accuracy simply is defined as the number of correct predictions divided by the total number of predictions there are [30]. This means that the lower the loss, and the higher the accuracy, the better the prediction model. The trials that we ran with the number of outcomes, number of rows, and length of the rows are listed in Figure 3.1.

To assess how the length of the rows affects the testing loss and accuracy, we can compare the results from trials 3 and 4, 5 and 6, and the results from

Trial Number	Number of Outcomes	Number of Rows	Length of Rows	Testing Loss	Accuracy
1	2	20000	250	0.10028	0.96949
2	2	10000	250	0.14859	0.958
3	2	800	250	0.45067	0.82916
4	2	800	60	1.32729	0.6875
5	3	126	60	0.00032	1
6	3	126	30	0.00115	1
7	5	300	60	1.83227	0.66666
8	5	300	30	0.90824	0.66666

Figure 3.1: Results of Overall Experiment

trials 7 and 8. In the first instance, when looking at trials 3 and 4, we see that as the length of the rows decreases from 250 to 60, the testing loss increases and the accuracy decreases. This means that the model's prediction gets worse. In the instance of trials 5 and 6, as the length of the rows decreases, the testing loss increases; however, the accuracy remains the same. This means that the model has a greater error as the length of the rows decreases but the same accuracy. In the example comparing 7 and 8, however, as the length of the rows decreases, the accuracy remains the same, and the loss actually decreases. However, it is important to note that in this case, the accuracy and the testing loss levels reveal that this model does not do a very good job at predicting.

To see the impact that the number of outcomes has on the ability of the model to predict, we can look at the results from trials 5 and 7, and 6 and 8. Although these pairs do not have the same number of rows, since we simply altered the example of 5 outcomes by taking away two of the existing pathobiomes, it results in fewer rows. Comparing trials 5 and 7, we see that the testing loss greatly increases as the number of outcomes increases, and the accuracy decreases as well. This is the same when comparing the results for trials 6 and 8.

To identify how the number of rows can affect the testing loss and accuracy of the model, we can compare trials 1, 2, and 3. As the number of rows decreases, the testing loss increases and the accuracy decreases. This means that more rows

Figure 3.2: Trial 5 Output

make for a more accurate prediction model.

To show what the outcome of our best trial looks like, Figure 3.2 shows the results from running trial 5.

We also wanted to do a small experiment to see if the accuracy or testing loss would improve if we ran the example with five outcomes but only had it predict two outcomes. Basically, we wanted to see if we used all the data from trial 7 but only had the model predict if it was salmonella or not. These were our results:

Testing loss: 0.2023201286792755, acc: 0.8461538553237915

This shows us that, in comparison to trial 7, the testing loss decreases and the accuracy increases, meaning with only two outcomes to choose from, the model makes better predictions.

Using our trial with the best outcome for testing loss and accuracy, we then decided to take it a step further to see if the batch size or the number of epochs in the model had any effect on the ability of the model to predict. For a machine learning model, the batch size is essentially how samples of the given data the model goes through before updating its internal parameters [31]. "Think of a batch as a for-loop iterating over one or more samples and making predictions. At the end of the batch, the predictions are compared to the expected output variables and an error is calculated" [31]. The number of epochs in machine learning is "... the number of times that the learning algorithm will work through the entire training dataset."[31].

Trial Number	Batch Size	Testing Loss	Accuracy
1	5	0.00021	1
2	10	0.00032	1
3	20	0.00265	1
4	40	0.00303	1
5	50	0.01026	1

Figure 3.3: Results of Batch Size Experiment

Trial Number	Epochs	Testing Loss	Accuracy
1	5	0.29089	1
2	25	0.00106	1
3	50	0.00083	1
4	100	0.00032	1
5	200	0.0002	1

Figure 3.4: Results of Epochs Experiment

Using trial 5, with 3 outcomes and 126 rows with a length 60, we ran an experiment to see the effect of increasing the batch size. In our original example, the batch size was set at 10. The trials that we ran are listed in Figure 3.3.

This outcome reveals that as the batch size increases, so does the testing loss. This means that the model makes better predictions when the batch size is smaller.

Using the same trial 3 from Figure 3.1, we then did a similar experiment with the number of epochs shown in Figure 3.4. This reveals that as the number of epochs increases, the model becomes better at predicting the outcome.

4 Conclusion

The results from the three experiments on the machine learning model have interesting implications for pathobiome detection. The first experiment shows that as the number of samples in the data increases, the model becomes more accurate and also has a lower testing loss. It is interesting that there does not seem to be a specific correlation between the length of the rows and the accuracy or testing loss. However, there does seem to be a correlation between the number of outcomes and the testing loss and accuracy as well. As the number of outcomes increases, so does the testing loss, and the accuracy decreases as well. The outcomes of our results for the batch size and number of epochs experiments were what we would have expected. As the number of epochs increased and as the batch size decreased, our model gave us a better prediction.

This means that in order for food safety decision-makers to have accurate and precise predictions when it comes to detecting pathobiomes, specifically that of salmonella, they will need DNA input data that has samples in the 1000s, not 100s. They will also need data that, if possible, is limited to a smaller number of outcomes to predict. Alternatively, as the number of classification outcomes increases, the number of samples required should increase linearly. This means it may be better for decision-makers to be deciding between only a couple pathobiomes to detect. Decision makers will also require a data model to run with the smallest batch size possible and as many epochs as possible. All of these things will help to give stakeholders the ability to detect pathobiomes more accurately in a machine-learning model.

Bibliography

[1] B. K, "5 essential steps for every deep learning model! — by bharath k — towards data science," https://towardsdatascience.com/5-essential-steps-for-every-deep-learning-model-30f0af3ccc37, November 2022, (Accessed on 03/04/2023).

[2] C. J. Grim, N. Daquigan, T. S. Lusk Pfefer, A. R. Ottesen, J. R. White, and K. G. Jarvis, "High-resolution microbiome profiling for detection and tracking of salmonella enterica," *Frontiers in microbiology*, vol. 8, p. 1587, 2017.

[3] C. E. Rainwater, "Improving food safety of pork supply chain, walmart foundation," 2021.

[4] "Top countries for pork production - source oecd," https://www.nationmaster.com/nmx/ranking/pork-production, (Accessed on 03/04/2023).

[5] "Agricultural output - meat consumption - oecd data," https://data.oecd.org/agroutput/meat-consumption.htm, (Accessed on 03/04/2023).

[6] "Foodborne diseases," https://www.who.int/health-topics/foodborne-diseases#tab=tab_2, (Accessed on 03/04/2023).

[7] Y. Li, Y. Huang, J. Yang, Z. Liu, Y. Li, X. Yao, B. Wei, Z. Tang, S. Chen, D. Liu *et al.*, "Bacteria and poisonous plants were the primary causative hazards of foodborne disease outbreak: a seven-year survey from guangxi, south china," *BMC public health*, vol. 18, no. 1, pp. 1–8, 2018.

[8] A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han, and L. Zhang, "Review on the application of machine learning algorithms in the sequence data mining of dna," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 1032, 2020.

[9] D. Bass, G. D. Stentiford, H.-C. Wang, B. Koskella, and C. R. Tyler, "The pathobiome in animal and plant diseases," *Trends in ecology & evolution*, vol. 34, no. 11, pp. 996–1008, 2019.

[10] "The 'pathobiome' – a new understanding of disease – sciencedaily," https://www.sciencedaily.com/releases/2019/09/190912113238.htm, (Accessed on 03/04/2023).

[11] N. Munck, P. M. K. Njage, P. Leekitcharoenphon, E. Litrup, and T. Hald, "Application of whole-genome sequences and machine learning in source attribution of salmonella typhimurium," *Risk Analysis*, vol. 40, no. 9, pp. 1693–1705, 2020.

[12] "Exploring a nucleotide sequence using the sequence viewer app - matlab & simulink," https://www.mathworks.com/help/bioinfo/ug/importing-viewing-and-exploring-a-nucleotide-sequence-using-a-graphical-interface.html, (Accessed on 03/13/2023).

[13] "Get started with bioinformatics toolbox," https://www.mathworks.com/help/bioinfo/getting-started-with-bioinformatics-toolbox.html, (Accessed on 03/13/2023).

[14] "What is an orf? — broad institute," https://www.broadinstitute.org/blog/what-orf, (Accessed on 03/13/2023).

[15] E. Wilson, "Modeling dna sequences with pytorch — by erin wilson — towards data science," https://towardsdatascience.com/modeling-dna-sequences-with-pytorch-de28b0a05036, September 2022, (Accessed on 03/04/2023).

[16] "deep_learning_dna/predict_seq.py at master · onceupon/deep_learning_dna · github," https://github.com/onceupon/deep_learning_DNA, (Accessed on 04/02/2023).

[17] "Convolution in one dimension for neural networks," https://e2eml.school/convolution_one_d.html#:~:text=A\%20convolution\%20layer\%20accepts\%20a,on\%20to\%20the\%20next\%20layer., (Accessed on 05/04/2023).

[18] "A gentle introduction to pooling layers for convolutional neural networks - machinelearningmastery.com," https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/#:~:text=Maximum\%20pooling\%2C\%20or\%20max\%20pooling,the\%20case\%20of\%20average\%20pooling., (Accessed on 05/04/2023).

[19] "The most intuitive and easiest guide for convolutional neural network — by jiwon jeong — towards data science," https://towardsdatascience.com/the-most-intuitive-and-easiest-guide-for-convolutional-neural-network-3607be47480, (Accessed on 05/04/2023).

[20] "A complete understanding of dense layers in neural networks," https://analyticsindiamag.com/

a-complete-understanding-of-dense-layers-in-neural-networks/, (Accessed on 05/04/2023).

[21] "A gentle introduction to dropout for regularizing deep neural networks - machinelearningmastery.com," https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/, (Accessed on 05/04/2023).

[22] S. Delannoy, S. Cadel-Six, L. Bonifait, M.-L. Tran, E. Cherchame, L. Baugé, K. Romero, S. Rouxel, A. Thépault, C. Cordevant *et al.*, "Closed genome sequence of a salmonella enterica serovar bovismorbificans strain isolated from dried pork sausage associated with an outbreak in france," *Microbiology Resource Announcements*, vol. 10, no. 40, pp. e00 662–21, 2021.

[23] "Acgt," https://www.genome.gov/genetics-glossary/acgt#:~:text=ACGT\%20is\%20an\%20acronym\%20for,and\%20cytosine\%20pairs\%20with\%20guanine, (Accessed on 05/04/2023).

[24] "Foodborne germs and illnesses — cdc," https://www.cdc.gov/foodsafety/foodborne-germs.html, (Accessed on 04/02/2023).

[25] A. Zamora-Vélez, J. Triviño, S. Cuadrado-Ríos, F. Lora-Suarez, and J. E. Gómez-Marín, "Detection and genotypes of toxoplasma gondii dna in feces of domestic cats in colombia," *Parasite*, vol. 27, 2020.

[26] S. Cai, D. Y. Kabuki, A. Y. Kuaye, T. G. Cargioli, M. S. Chung, R. Nielsen, and M. Wiedmann, "Rational design of dna sequence-based strategies for subtyping listeria monocytogenes," *Journal of Clinical Microbiology*, vol. 40, no. 9, pp. 3319–3325, 2002.

[27] S. Kundu, J. Lockwood, D. P. Depledge, Y. Chaudhry, A. Aston, K. Rao, J. C. Hartley, I. Goodfellow, and J. Breuer, "Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients," *Clinical infectious diseases*, vol. 57, no. 3, pp. 407–414, 2013.

[28] G. Zirnstein, L. Helsel, Y. Li, B. Swaminathan, and J. Besser, "Characterization of gyra mutations associated with fluoroquinolone resistance in campylobacter coli by dna sequence analysis and mama pcr," *FEMS Microbiology Letters*, vol. 190, no. 1, pp. 1–7, 2000.

[29] "Descending into ml: Training and loss — machine learning — google developers," https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss#:~:text=Loss\%20is\%20the\%20penalty\%20for,otherwise\%2C\%20the\%20loss\%20is\%20greater, (Accessed on 05/01/2023).

[30] "Classification: Accuracy — machine learning — google developers," https://developers.google.com/machine-learning/crash-course/classification/accuracy#:~:text=Accuracy\%20is\%20one\%20metric\%20for,predictions\%20Total\%20number\%20of\%20predictions, (Accessed on 05/01/2023).

[31] "Difference between a batch and an epoch in a neural network - machinelearningmastery.com," https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/, (Accessed on 05/01/2023).