DISCOVERY OF RELATION BETWEEN *Listeria* AND OTHER BACTERIAL CONTAMINATION USING CLASSIFICATION REFINEMENT TECHNIQUE

Miss Napas Jeamchotpatanakul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science and Information Technology
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2017

การค้นพบของความสัมพันธ์ระหว่างการปนเปื้อนเชื้อลิสทีเรียกับแบคทีเรียอื่นๆ โดยใช้เทคนิคการจำแนกประเภทอย่างละเอียด

นางสาวนภัส เจียมโชติพัฒนกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

Thesis Title	DISCOVERY OF RELATION BETWEEN *Listeria* AND OTHER BACTERIAL CONTAMINATION USING CLASSIFICATION REFINEMENT TECHNIQUE
By	Miss Napas Jeamchotpatanakul
Field of Study	Computer Science and Information Technology
Thesis Advisor	Assistant Professor Saranya Maneeroj, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

....Dean of the Faculty of Science

(Professor Polkit Sangvanich, Ph.D.)

THESIS COMMITTEE

....Chairman

(Associate Professor Peraphon Sophatsathit, Ph.D.)

....Thesis Advisor

(Assistant Professor Saranya Maneeroj, Ph.D.)

....External Examiner

(Assistant Professor Saichon Jaiyen, Ph.D.)

นภัส เจียมโชติพัฒนกุล : การค้นพบของความสัมพันธ์ระหว่างการปนเปื้อนเชื้อลิสทีเรียกับแบคทีเรียอื่นๆ โดยใช้เทคนิคการจำแนกประเภทอย่างละเอียด (DISCOVERY OF RELATION BETWEEN *Listeria* AND OTHER BACTERIAL CONTAMINATION USING CLASSIFICATION REFINEMENT TECHNIQUE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. ศรันญา มณีโรจน์, หน้า.

การบริโภคอาหารโดยมีเนื้อไก่ส่วนประกอบเป็นที่แพร่หลายสำหรับมนุษย์เพราะฉนั้นการผลิตและส่งออกเนื้อไก่จึงเป็นสิ่งสำคัญสำหรับหลายประเทศ โดยเฉพาะประเทศไทยที่ผลอันดับจำนวนการส่งออกเนื้อไก่เป็นอันดับที่ 4 ของโลก ซึ่งในการส่งออกเนื้อไก่นั้นจะมีระบบและมาตราฐานควบคุมซึ่งในแต่ละประเทศก็จะมีระเบียบบางอย่างที่แตกต่างกันออกไป แต่มาตราฐานที่ทุกประเทศตระหนักถึงคือการตรวจพบเชื้อลิสทีเรียในเนื้อไก่ ถ้าหากมีการตรวจพบเชื้อลิสทีเรียในเนื้อไก่ จะมีผลกระทบต่อประเทศผู้ส่งออกนั้นๆ สาเหตุที่ทุกประเทศตระหนักถึงความรุนแรงของเชื้อลิสทีเรียเนื่องจากเชื้อดังกล่าวสามารถส่งผลต่อชีวิตของมนุษย์โดยเฉพาะหญิงตั้งครรภ์และทารกที่อยู่ในครรภ์ ดังนั้นนักวิทยาศาสตร์ทางอาหารจึงมีความสนใจที่จะป้องกันการเกิดเชื้อลิสทีเรีย ซึ่งวิธีในปัจจุบันคือการทดสอบจุลชีววิทยา วิธีดังกล่าวนั้นยังไม่สามารถป้องกันการเกิดลิสทีเรียได้

ยังไม่มีนักวิจัยกลุ่มใดประยุกต์การทดสอบจุลชีววิทยาเข้ากับแนวคิดของคอมพิวเตอร์ จากการทดสอบจุลชีววิทยาทำให้มีชุดของผลข้อมูลเป็นจำนวนมากพอที่จะประยุกต์ใช้กับแนวคิดของคอมพิวเตอร์ ดังนั้นงานวิจัยนี้จึงเสนอการค้นพบปัจจัยที่ทำให้เกิดเชื้อลิสทีเรียโดยใช้เทคนิคการจำแนกประเภทอย่างละเอียดประกอบด้วย random forest and linear regression ที่จะใช้สำหรับการทำนาย และ Naïve Bayes สำหรับการประเมินความถูกต้องของแนวคิดที่นำเสนอ ซึ่งแนวคิดดังกล่าวสามารถบ่งชี้ให้เห็นถึงแบคทีเรียที่มีความเกี่ยวข้องต่อการปนเปื้อนเชื้อลิสทีเรียและจำนวนของแบคทีเรียดังกล่าวที่จะก่อให้เกิดการปนเปื้อนของเชื้อลิสทีเรีย

ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	ลายมือชื่อนิสิต	....
		ลายมือชื่อ อ.ที่ปรึกษาหลัก	....
สาขาวิชา	วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ		
ปีการศึกษา	2560		

5972632723 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
KEYWORDS: CLASSIFICATION REFINEMENT / RANDOM FOREST / LINEAR REGRESSION / NAIVE BAYES

NAPAS JEAMCHOTPATANAKUL: DISCOVERY OF RELATION BETWEEN *Listeria* AND OTHER BACTERIAL CONTAMINATION USING CLASSIFICATION REFINEMENT TECHNIQUE. ADVISOR: ASST. PROF. SARANYA MANEEROJ, Ph.D., pp.

People often use chicken to prepare their diary meal which creates a great demand for food producers and exporters in Thai food industry. Thailand is ranked fourth in chicken exporters in the world. Standards on chicken meat regulations differ from country to country. One commonality remains: controlling the bacteria that can affect human life, particularly pregnant women and the unborn. One of the important bacteria is called *Listeria* which receives high attention for the industry to prevent their contamination in chicken. A microbiological test is usually conducted to analyze the data from industry. However, the microbiological test is incapable of identifying other variants from Listeria contamination. There is no computer model that can be used in Listeria contamination analysis. Anyhow, such tests create plenty of data that can be applied with a computer model to investigate the factors that causes Listeria contamination.

This research proposes a classification refinement technique which composes of random forest and linear regression to predict the variants of Listeria contamination. Moreover, Naïve Bayes is used to assess the model correctness. This proposed procedure can reveal the important causes relating to variants and the conditions on the number of these variants that produce Listeria contamination.

Department:	Mathematics and Computer Science	Student's Signature	....
		Advisor's Signature	....
Field of Study:	Computer Science and Information Technology		
Academic Year:	2017		

ACKNOWLEDGEMENTS

I would like to thank all the people who have a part in completed this thesis.

First of all, I am very much obliged and grateful to my research advisor, Assistant Professor Dr. Saranya Maneeroj for useful advice and valuable suggested direction to make me completed this thesis including the guideline of classification methodology, the structure of proposal and thesis and the presentation practicing.

The second thank, I am so grateful to food scientist who gives me the particular information in food domain including dataset, the current used methodology and the current problems.

Thirdly, I would say thank you to program chair, Associate Professor Dr.Peraphon Sophatsathit and external examiner, Assistant Professor Dr.Saichon Jaiyen for the suggestions and great opinion.

The final thank, I would give a big thank to my parents and family who are always support me for everything.

CONTENTS

Page

Page

Page

Content of Tables

Content of Figures

Chapter 1.
INTRODUCTION

This chapter composes of 6 main sections. The first section is Background and Importance which describe the overview of the problems and current solutions. The second section introduces the objective of this work. The third section describes the scope and the limitations of this work. The fourth section is the main problem and the fifth section describes the expected outcome of this work. The last section introduces the overall structure of the rest of this document.

1.1 Background and importance

There are many of countries that produce food and export food. Thailand is the one of largest food producers in the world. Many food products are produced in Thailand such as chicken, canned tuna, canned pineapple, rice, and frozen seafood. Thus, Thai Ready-to-Eat food gains high demands of overseas exporting as Thailand has good reputation in quality, taste, and nutrition. Thailand is the fourth largest exporting in cooked chicken in the world. Two principal cooked chicken markets of Thailand are Japan and EU[1] who impose high standards for cooked chicken meat imports. Some problems arise in the processing of food products that have some bacteria contamination. It can occur on any procedures along the processing line and storage. One of the important bacteria is called *Listeria*. *Listeria* can be found in any equipment including corner of mechanism, drain, floor, and others. *Listeria* is hard to spot and prevent. So, the important indicator to prevent *Listeria* contamination include *TVC*, *E.coli*, *Coliform,* and *Enterococci*. Despite some regulatory standards of the Department of Livestock Development of Thailand, various countries including Japan and the EU who import ready-to-eat meat and poultry products require a zero tolerance (negative in 25 g sample for L. Monocytogenes)[2]. Hence, laboratory test is an important step to prove that such cooked chicken is safe. This instigates many researchers to study on disposing and preventing *Listeria* contamination. The recent preventing *Listeria* contamination methodology from food scientists has only studied on microbiological test and adapted it to basic statistics. It means that they manually predict the occurrence of *Listeria* contamination by using microbiological test and analysis by expert food scientists in laboratory. These examination and analysis only represent the result after the *Listeria* has already contaminated the product. Thus, the recent method failed investigate what factors caused *Listeria* contamination in cooked chicken meat. However, there are many data from microbiological tests that can be used to analyze and predict what relate the bacteria to *Listeria* contamination with the help of a computer model. From these data, bacteria such as *TVC*, *E.coli*, *Coliform,* and *Enterococci* can be used as the features in classification technique.

Some computer researchers applied classification method to food safety. A group of researchers take some characteristics of SVM method which can reinforce the massive number of record data and to find the bacteria and the results of this research with good speed and accuracy[3]. Meanwhile, another group of researchers enhance the predictive model in food safety area by using decision analysis that includes balanced iterative reducing and clustering hierarchies for data mining algorithm[4].

It is not only food safety area that applies the computer model but also bacteria detection classification method. For example, SVM is used to analyze the *E.coli* contamination by employing the characteristic of *E.coli* as the features[5]. On the other hand, random forest is used to analyze the tuberculosis bacteria using the characteristic of tuberculosis as the features[6].

Apparently, the major research in food area applied only single classification method. It could be exceeded to embed the classification methods together. That might be effectiveness to the result and get the exact important variants that cause contamination of bacteria.

This study aims to present the potential of the classification refinement methodology which is helpful to predict the bacterial factor of *Listeria* contamination. There are four significant bacterial factor including *E.coli*, *Coliform*, *Enterococci*, and *TVC*. Also, the classification refinement technique composes of two main classification methods which include random forest and linear respectively. Furthermore, to reassure the model, Naïve Bayes is used. The datasets for this work consists of two type of dataset which are cooked chicken and cooked rice. Both datasets return the same direction result. The classification refinement methodology is applied to the cooked chicken dataset and the result returns with superb result. Then, model reassuring is the next step so cooked rice dataset is applied on the proposed model. This dataset includes *E.coli*, *Coliform* and *TVC* for features. The result indicates that *E.coli* is a main important factor to detect *Listeria* contamination to both datasets.

1.2 Objective

To find the related bacteria factors which lead to *Listeria* contamination using classification refinement techniques i.e., random forest and Linear regression. Additionally, Naïve Bayes is applied to assess the correctness of the model.

1.3 Scope of thesis and constraints

There are two sample datasets from microbiological testing. They are used with the limited details as follows:

1.3.1 Cooked chicken dataset

There are 2375 records including 150 positive classes and 2,225 negative classes. Also, the features consist of four main features which are *TVC*, *E.coli*, *Coliform,* and *Enterococci*.

1.3.2 Cooked rice dataset

There are 159 records including 17 positive classes and 142 negative classes. The features consist of four main features which are *TVC*, *E.coli*, and *Coliform*.

1.3.3 Dataset limitations

The result of proposed method specifies how the dataset is gathered and the type of product in the dataset. However, concepts of the model can be applied to find other related factors of bacteria in other microbiological tests.

1.4 Problem formulation

In food safety area, *Listeria* contamination is necessary to research and prevent. The current method to detect *Listeria* contamination is microbiological testing manually. Then, the result tests are analyzed by the food scientists. Based on the current method, it failed to predict the related factor to *Listeria* contamination. It means that there is no method to prevent causing *Listeria* contamination. In part of prediction model, there is also none of research for *Listeria* contamination that using computer model. Indeed, there are some researches which use computer model to find other bacteria by using classification techniques. However, they used only single classification model individually by reason of complication of combination the classification model. Though, it may improve the accuracy of the model, if it is able to combine the classification model appropriately.

1.5 Expected outcome

This research aims to consign the classification refinement technique which can help find the related bacteria factors causing *Listeria* contamination. The outcome can represent the important indicators which identify *Listeria* contamination. The result expresses the conditions of bacteria unit which lead to *Listeria* contamination.

1.6 Thesis structure

The rest of this research is organized as follows. Chapter 2. describes literature review. Chapter 3. describes Research methodology. Chapter 4 presents evaluation methodology. Discussion and conclusion are given in Chapter 5.

Chapter 2.
Literature Review

This chapter presents classification methods and the background of these classification methods in food domain. There are 4 main sections which start from random forest concept, followed by linear regression concept. The rest of the two sections are used classification methods in food safety and bacteria detection.

2.1 Random forest

Random forest is the one of accompany learning algorithm for classification and regression method. Random forest concept uses some basic approach from decision tree. In another word, multiple decision trees are generated and gathering to be the element of random forest. It may affect to the power of accuracy and prediction result. According to the concept of random forest is similar to decision tree, the first step is extracted features. Then, those features are randomly selected to construct each decision trees in logically. There are some methods to select the features including Gini index, information gain and gain ratio. Also, the random forest has capability of feature importance which can score each feature after applied and measure the result. It decreases the impurity beyond the trees in forest. After random trees are generated, all those result trees are voted to get the final result.

2.2 Linear regression

Linear regression is the one of classification method for prediction in a way of statistical model. Linear regression analyses the whole dataset and return the result in equation to plot the graph. Thus, the result of linear regression represents the relationship between the variables. Also, the coefficient in each variables can instruct the weight how is the importance of the variable.

2.3 Food Safety

Food Safety is the process that controls, arrange, store the food. Thus, such process is able to reduce risk from food that may affect consumers for instance sickness and disease. In consideration, the bacteria are able to be attached into food and consumers can infect those bacteria. It may affect to human even more or less. So, food safety is a general attention that may the reason to many researchers are concerned to find the solution to that problem as follow:

2.3.1 Food safety based on SVM concept

There are a group of researchers who work on SVM, CASCADE SVM AND CLUSTER SVM to analyze dairy products in food safety area. The capability of SVM can support running massive dataset. In contrast, SVM model takes a plenty of time to run and the result is not good enough. That is the reason why they take cluster SVM and cascade SVM which based on SVM concept. To ensure that cascade SVM which based on SVM return the better result than single SVM, the researcher will compare among such three models. They use the same dataset which is dairy product to all methods as Figure 1.

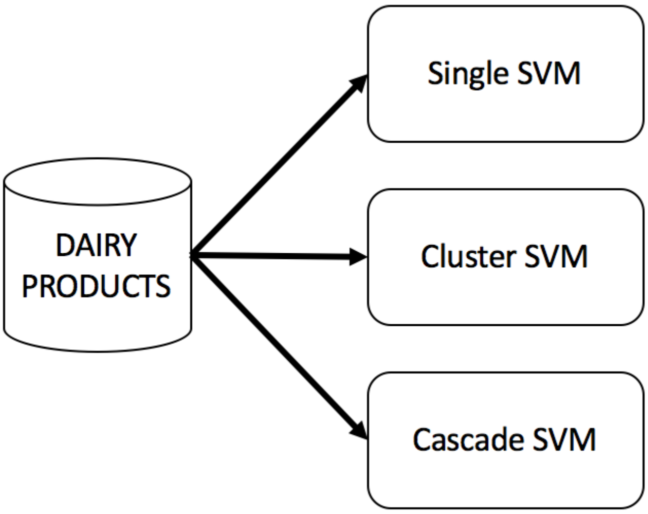


Figure 1. The overview model of the first food safety research using classification method

For the SVM is applied and then the time taken and memory usage are too much. Thus, the cluster SVM and cascade SVM are used to solve those problem by using data decomposition thinking. The cluster SVM will separate data into N blocks. For the first layer will return the support vector result in each n blocks. Then the non-support vectors in the training results are removed from each small data set. Next, a half of N blocks are merged to be the data block in second layer and consecutively combine the rest two training data until getting the last result in third layer as Figure 2. All three layers will be repeated until getting the optimal result.

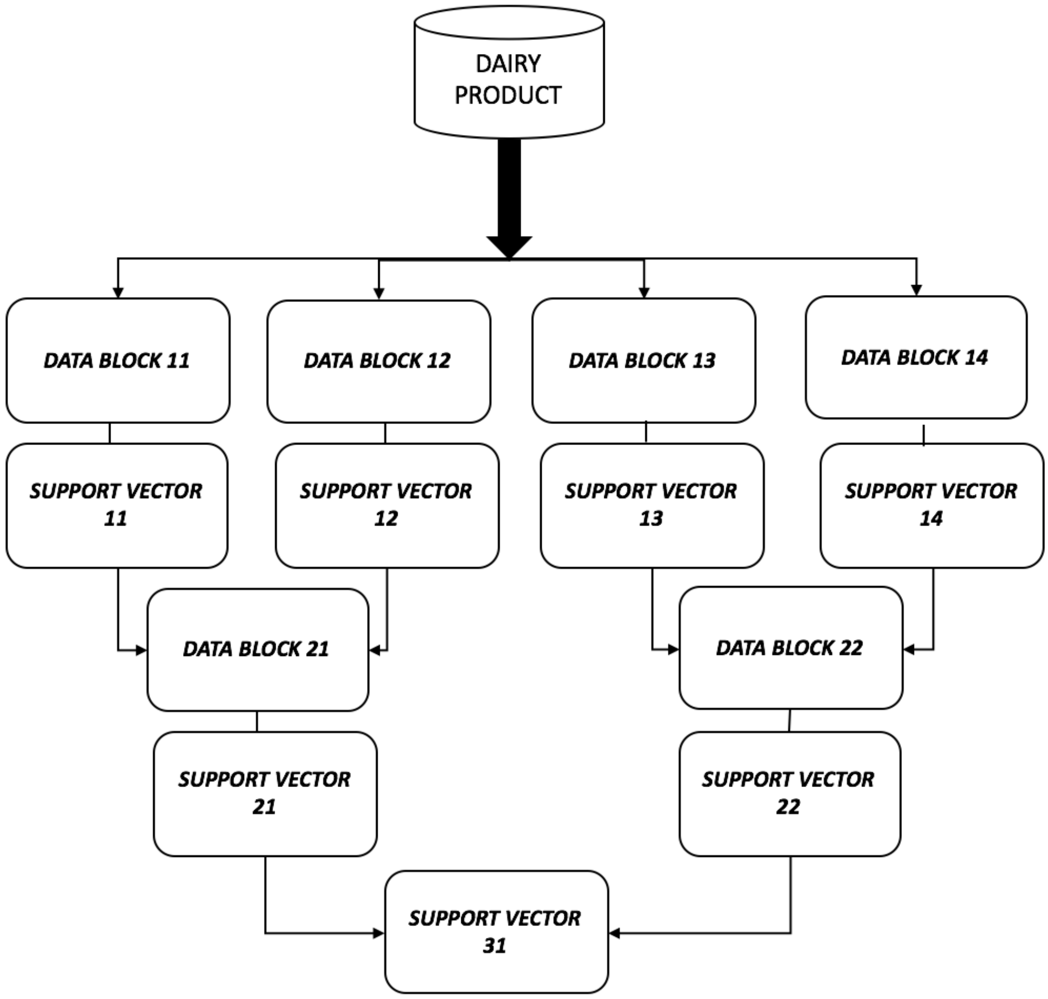


Figure 2. The cluster SVM model concept

According to the limitation of the number of block in cluster SVM, it may cause to the enormous data to separate in each blocks which makes a dense data blocks. Thus, the researchers propose the cascade SVM which based on cluster SVM but it is able to separate more than the number of blocks in the cluster SVM concept. Also, they take MapReduce concept to enhance merge algorithm. This concept can reduce the time taken and memory usage from the single SVM and also get the better result [3].

2.3.2 Data analysis on food safety and traceability system

Another concept to apply the computer model with food safety area is food safety and traceability system. The computer model can support in part of data mining and analysis. The system used Balanced Iterative Reducing and Clustering using Hierarchies concept which another called is dendrogram. There are two meaningful steps as shown in Figure 3.

The first step is scan database by using Linear-Bush-Tree for immune genetic algorithm, this algorithm can cope with huge data gathering and takes good speed of querying. The second step is clustering feature tree building which is dendrogram. Eventually, it returns the clustering results which represent two class of the quality of the product including faulty product or perfect product[4].

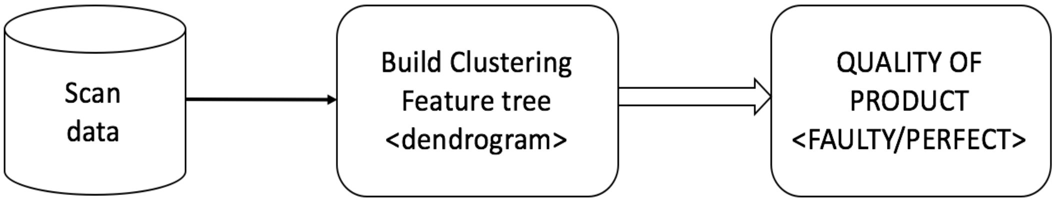

Scan data
Build Clustering Feature tree <dendrogram>
QUALITY OF PRODUCT <FAULTY/PERFECT>

Figure 3. The second model process for food safety using classification method

2.4 Bacteria Detection

Normally, bacteria detection learns the particular characteristics of each bacterium. Thus, to adapt with computer model, the image processing is mostly used to distribute the characteristics of each bacterium to be the feature in classification technique.

2.4.1 E.*coli* detection based on SVM

One of the researcher group in bacteria detection domain used SVM to breakdown the class of *E.coli* by learning the characteristic of this bacteria. There are three main steps of this model as shown in Figure 4. Firstly, *E.coli* are captured to be the image from biological microscope. The second step is extracting the feature from the specific character of *E.coli* for example area, perimeter, roundness, long-axis, short-axis and ellipticity. Finally, SVM is applied to detect *E.coli* and count the unit of *E.coli*. The result represents the distinguish class and quantity of *E.coli[5]*.

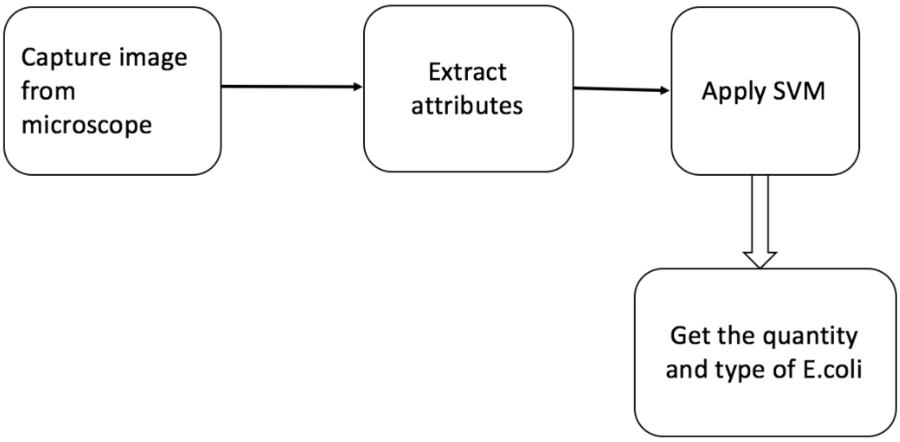

Capture image from microscope
Extract attributes
Apply SVM
Get the quantity and type of E.coli

Figure 4. The first model process for bacteria detection using classification method

2.4.2 *Tuberculosis* detection on fluorescent microscope system based on random forest

Another model for bacteria detection is for automation of *tuberculosis* detection on fluorescent microscope system. The researchers use the characteristics of random forest to apply with this system properly. There are 4 significant steps as shown in Figure 5.

The first step is captured *tuberculosis* image on fluorescent microscope which represents the bacteria on the green color. The second step is to erase the background of the image and keep only the element of bacilli in the image. The third step is extracted feature from such image including Hu moment invariants, geometric and binary shape properties, and histograms of oriented gradients. Finally, all data is applied with 3 classification methods which composes of random forest, linear SVM and cross-validation SVM. As the classification result, random forest returns the greatest accuracy result to compare with the rest of two classification methods [6].

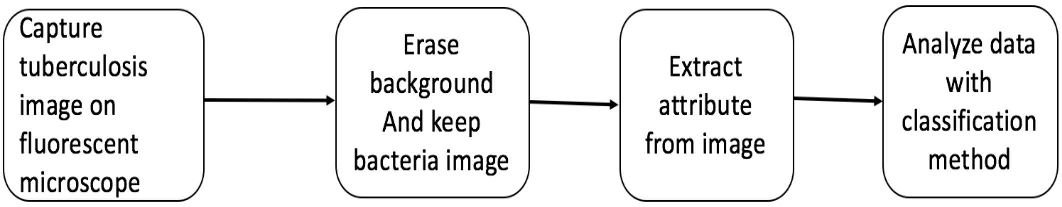

Capture tuberculosis image on fluorescent microscope
Erase background And keep bacteria image
Extract attribute from image
Analyze data with classification method

Figure 5. The model process for second bacteria detection using classification method

Chapter 3.
RESEARCH METHODOLOGY

According to enhance finding *Listeria* contamination model, classification refinement method is presented which includes random forest and linear regression to discover the *Listeria* contamination using computer model. Moreover, Naïve Bayes is used to assess the model. Cooked chicken and cooked rice are the datasets being applied to three methods. There are four main steps for this model as shown in Figure 6.

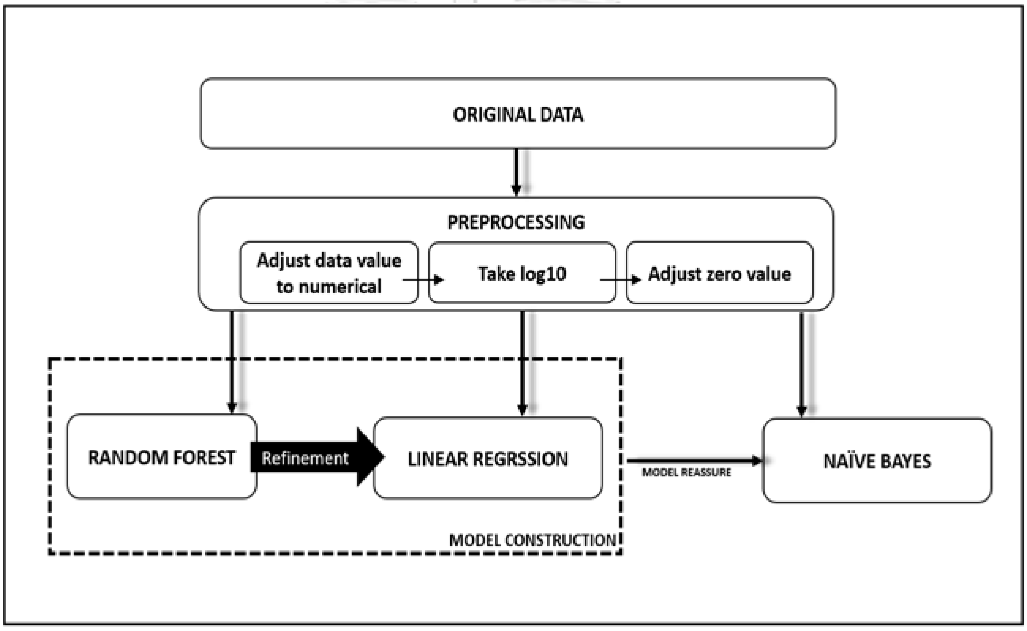


Figure 6. The proposed model concepts

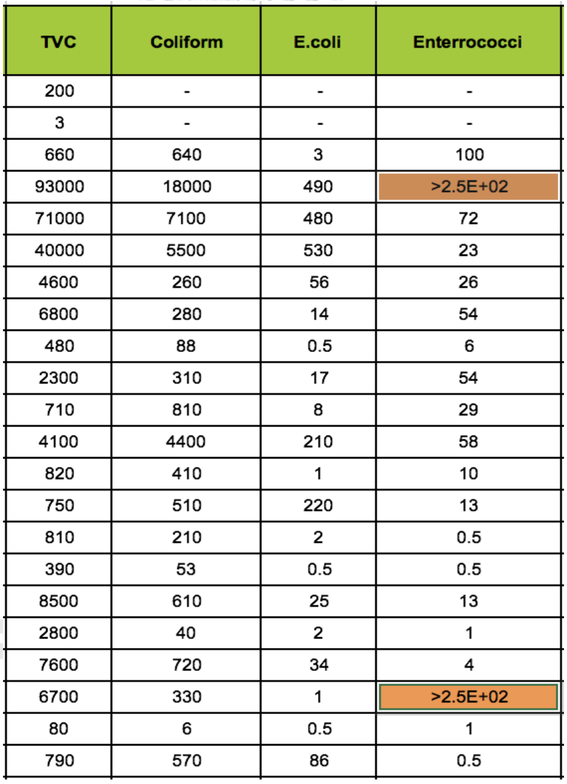
3.1 Preprocessing dataset

Based on the data from microbiological test, there are plenty of data in each batch test results. Each batch contains test result of several bacteria that can appear in food. They are used as the dataset. This model used two datasets which compose of cooked chicken and cooked rice. Cooked chicken dataset includes four main features which are *TVC*, *Coliform*, *E.coli,* and *Enterococci*. For cooked rice, there are three main features which include *TVC*, *Coliform,* and *E.coli*. They are slightly different but both have positive class (founded *Listeria*) and negative class (not founded *Listeria*).

The original data from laboratory are inappropriate to work on due to problems as follows.

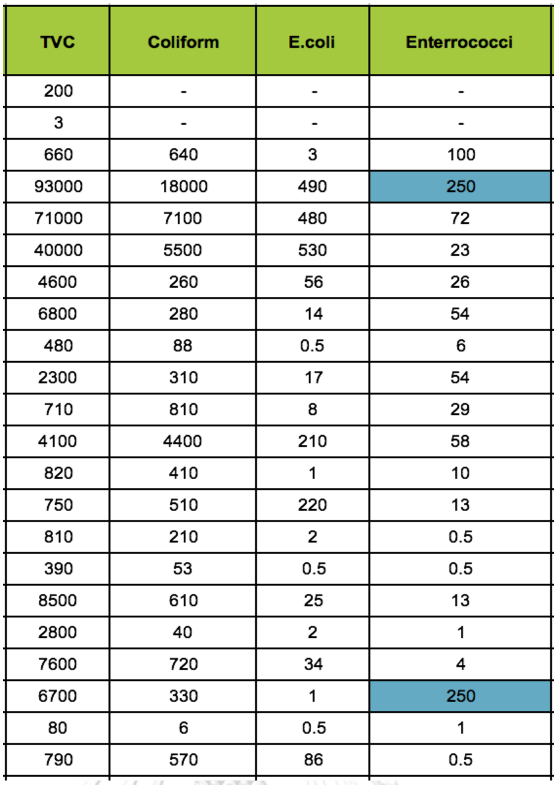
3.1.1 Non-numerical value

Some values of the data are non-numerical such as characters and abbreviations as shown in Figure 7. Hence, the non-numerical data are modulated from scientific notation to expressible as shown in Figure 8.



TVC	Coliform	E.coli	Enterrococci
200	-	-	-
3	-	-	-
660	640	3	100
93000	18000	490	>2.5E+02
71000	7100	480	72
40000	5500	530	23
4600	260	56	26
6800	280	14	54
480	88	0.5	6
2300	310	17	54
710	810	8	29
4100	4400	210	58
820	410	1	10
750	510	220	13
810	210	2	0.5
390	53	0.5	0.5
8500	610	25	13
2800	40	2	1
7600	720	34	4
6700	330	1	>2.5E+02
80	6	0.5	1
790	570	86	0.5

Figure 7. The example of non-numerical value in cooked chicken dataset

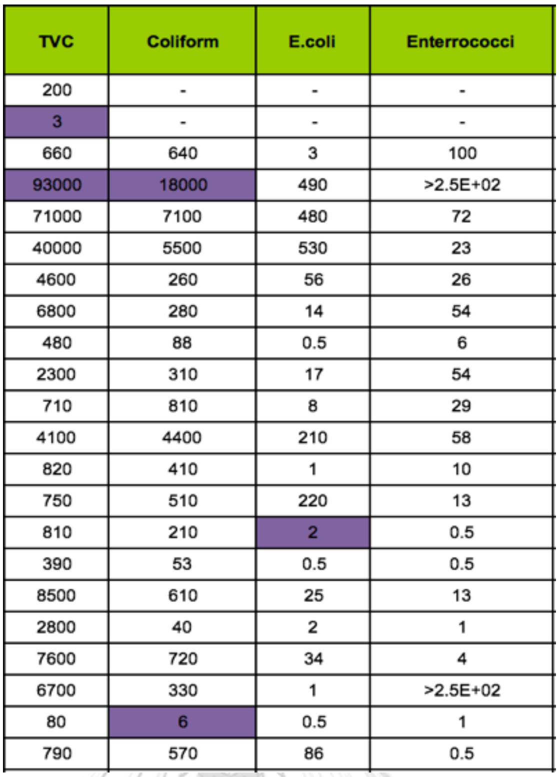


TVC	Coliform	E.coli	Enterrococci
200	-	-	-
3	-	-	-
660	640	3	100
93000	18000	490	250
71000	7100	480	72
40000	5500	530	23
4600	260	56	26
6800	280	14	54
480	88	0.5	6
2300	310	17	54
710	810	8	29
4100	4400	210	58
820	410	1	10
750	510	220	13
810	210	2	0.5
390	53	0.5	0.5
8500	610	25	13
2800	40	2	1
7600	720	34	4
6700	330	1	250
80	6	0.5	1
790	570	86	0.5

Figure 8. The result after adjusted the non-numerical to numerical

3.1.2 Wide range of value

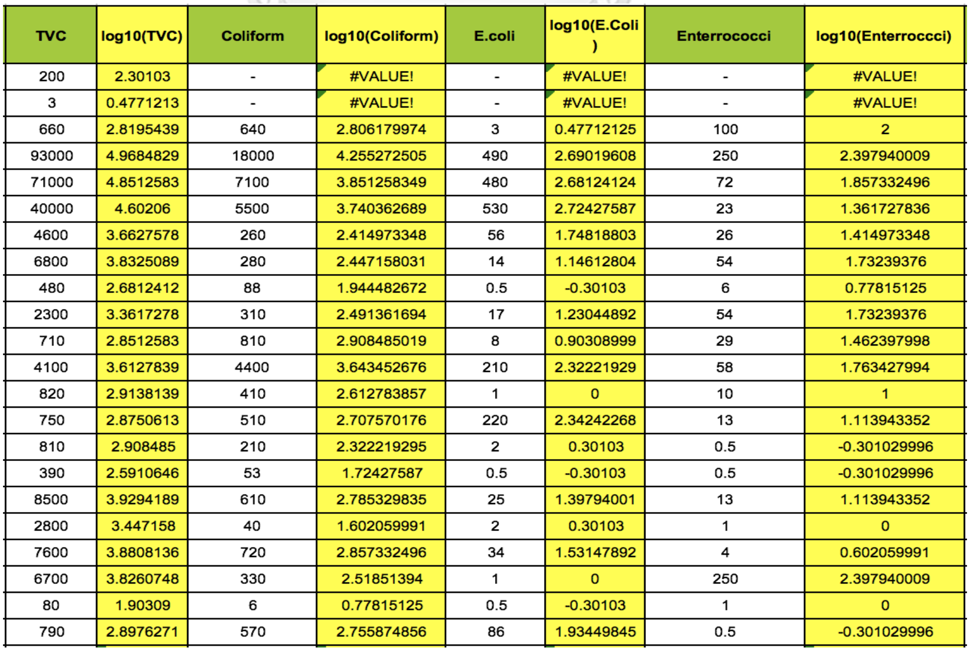
The next problem is about range of data value which are widely dispersed as shown in Figure 9. Examples in cooked chicken dataset, the value of *TVC* ranges from 3 to 93000, *Coliform* from 8 to 18000, *E.coli* from 0.5 to 530 and *Enterococci* from 0.5 to 250.



TVC	Coliform	E.coli	Enterrococci
200	-	-	-
3	-	-	-
660	640	3	100
93000	18000	490	>2.5E+02
71000	7100	480	72
40000	5500	530	23
4600	260	56	26
6800	280	14	54
480	88	0.5	6
2300	310	17	54
710	810	8	29
4100	4400	210	58
820	410	1	10
750	510	220	13
810	210	2	0.5
390	53	0.5	0.5
8500	610	25	13
2800	40	2	1
7600	720	34	4
6700	330	1	>2.5E+02
80	6	0.5	1
790	570	86	0.5

Figure 9. The example of the wide range value in cooked chicken dataset

To solve this problem, log 10 is used to reduce the data gap as shown in Figure 10.

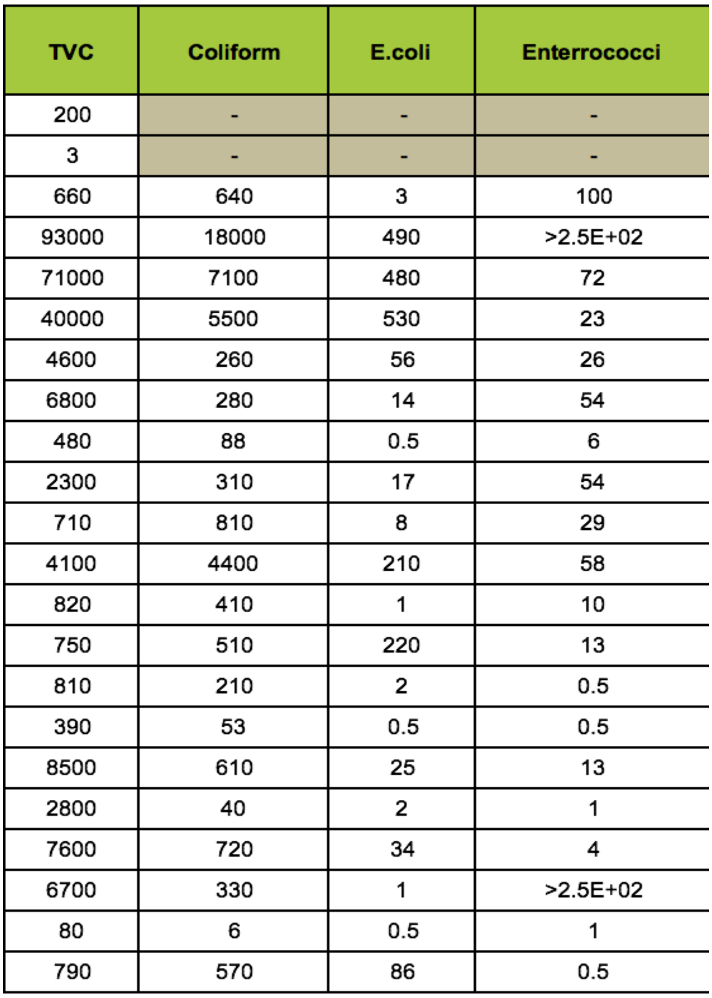


TVC	log10(TVC)	Coliform	log10(Coliform)	E.coli	log10(E.Coli)	Enterrococci	log10(Enterroccci)
200	2.30103	-	#VALUE!	-	#VALUE!	-	#VALUE!
3	0.4771213	-	#VALUE!	-	#VALUE!	-	#VALUE!
660	2.8195439	640	2.806179974	3	0.47712125	100	2
93000	4.9684829	18000	4.255272505	490	2.69019608	250	2.397940009
71000	4.8512583	7100	3.851258349	480	2.68124124	72	1.857332496
40000	4.60206	5500	3.740362689	530	2.72427587	23	1.361727836
4600	3.6627578	260	2.414973348	56	1.74818803	26	1.414973348
6800	3.8325089	280	2.447158031	14	1.14612804	54	1.73239376
480	2.6812412	88	1.944482672	0.5	-0.30103	6	0.77815125
2300	3.3617278	310	2.491361694	17	1.23044892	54	1.73239376
710	2.8512583	810	2.908485019	8	0.90308999	29	1.462397998
4100	3.6127839	4400	3.643452676	210	2.32221929	58	1.763427994
820	2.9138139	410	2.612783857	1	0	10	1
750	2.8750613	510	2.707570176	220	2.34242268	13	1.113943352
810	2.908485	210	2.322219295	2	0.30103	0.5	-0.301029996
390	2.5910646	53	1.72427587	0.5	-0.30103	0.5	-0.301029996
8500	3.9294189	610	2.785329835	25	1.39794001	13	1.113943352
2800	3.447158	40	1.602059991	2	0.30103	1	0
7600	3.8808136	720	2.857332496	34	1.53147892	4	0.602059991
6700	3.8260748	330	2.51851394	1	0	250	2.397940009
80	1.90309	6	0.77815125	0.5	-0.30103	1	0
790	2.8976271	570	2.755874856	86	1.93449845	0.5	-0.301029996

Figure 10. The result after adjusted the gap of value range by taking log 10

3.1.3 Missing value

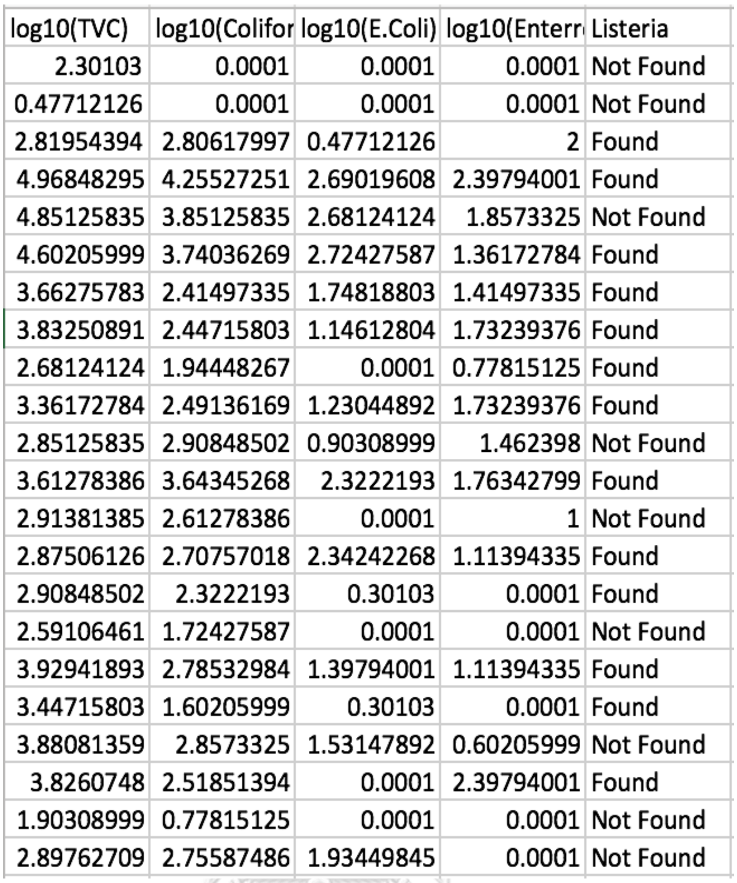
The third problem is about many missing values and zero data value which affect to analysis of data as shown in Figure 11. The solution is adjusting to the nearest zero that is not around the existing data value.



TVC	Coliform	E.coli	Enterrococci
200	-	-	-
3	-	-	-
660	640	3	100
93000	18000	490	>2.5E+02
71000	7100	480	72
40000	5500	530	23
4600	260	56	26
6800	280	14	54
480	88	0.5	6
2300	310	17	54
710	810	8	29
4100	4400	210	58
820	410	1	10
750	510	220	13
810	210	2	0.5
390	53	0.5	0.5
8500	610	25	13
2800	40	2	1
7600	720	34	4
6700	330	1	>2.5E+02
80	6	0.5	1
790	570	86	0.5

Figure 11. The example of missing value in cooked chicken dataset

After adjusting the zero value and missing value, the result is shown in Figure 12.



log10(TVC)	log10(Colifor	log10(E.Coli)	log10(Enterr	Listeria
2.30103	0.0001	0.0001	0.0001	Not Found
0.47712126	0.0001	0.0001	0.0001	Not Found
2.81954394	2.80617997	0.47712126	2	Found
4.96848295	4.25527251	2.69019608	2.39794001	Found
4.85125835	3.85125835	2.68124124	1.8573325	Not Found
4.60205999	3.74036269	2.72427587	1.36172784	Found
3.66275783	2.41497335	1.74818803	1.41497335	Found
3.83250891	2.44715803	1.14612804	1.73239376	Found
2.68124124	1.94448267	0.0001	0.77815125	Found
3.36172784	2.49136169	1.23044892	1.73239376	Found
2.85125835	2.90848502	0.90308999	1.462398	Not Found
3.61278386	3.64345268	2.3222193	1.76342799	Found
2.91381385	2.61278386	0.0001	1	Not Found
2.87506126	2.70757018	2.34242268	1.11394335	Found
2.90848502	2.3222193	0.30103	0.0001	Found
2.59106461	1.72427587	0.0001	0.0001	Not Found
3.92941893	2.78532984	1.39794001	1.11394335	Found
3.44715803	1.60205999	0.30103	0.0001	Found
3.88081359	2.8573325	1.53147892	0.60205999	Not Found
3.8260748	2.51851394	0.0001	2.39794001	Found
1.90308999	0.77815125	0.0001	0.0001	Not Found
2.89762709	2.75587486	1.93449845	0.0001	Not Found

Figure 12. The result after adjusted the missing values

3.2 Random forest

After solving the problems to complete the dataset, random forest method is used to extract some significant feature. There are three steps as follows.

3.2.1 Initial setting

The initial setting selects features as gain ratio, max depth as 20 and number as tree as 15.

3.2.2 Tree generating

Fifteen trees are generated. Each tree includes the conditions which indicate the important features that lead to *Listeria* contamination as shown in Figure 13 to Figure 27 below.

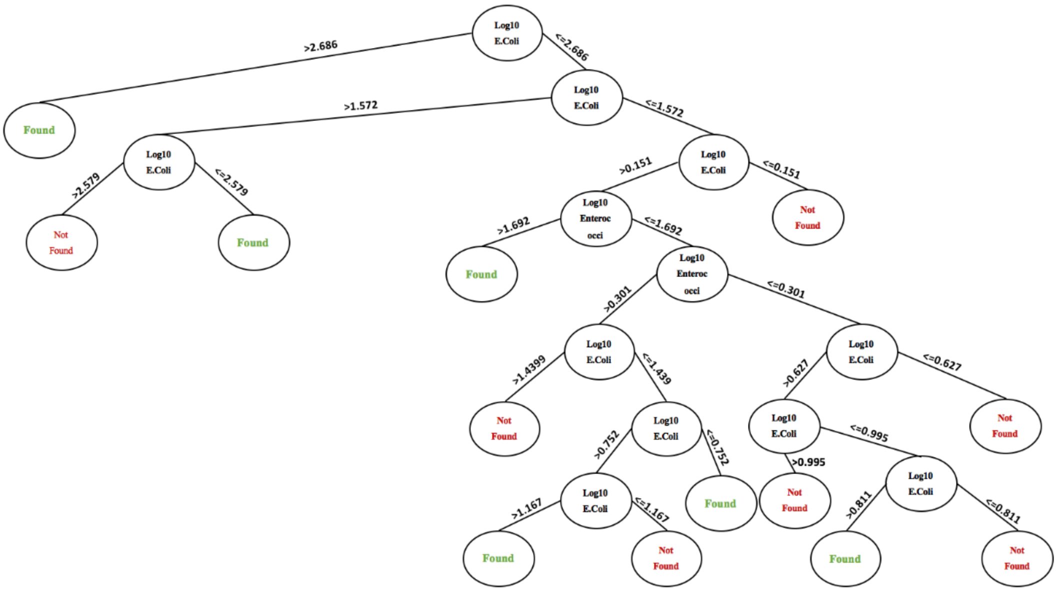


Figure 13. The first generated tree in cooked chicken dataset

For the first generated tree in cooked chicken dataset, there are 2 features, namely, *E.coli* and *Enterococci*, where *E.coli* is the outstanding feature.

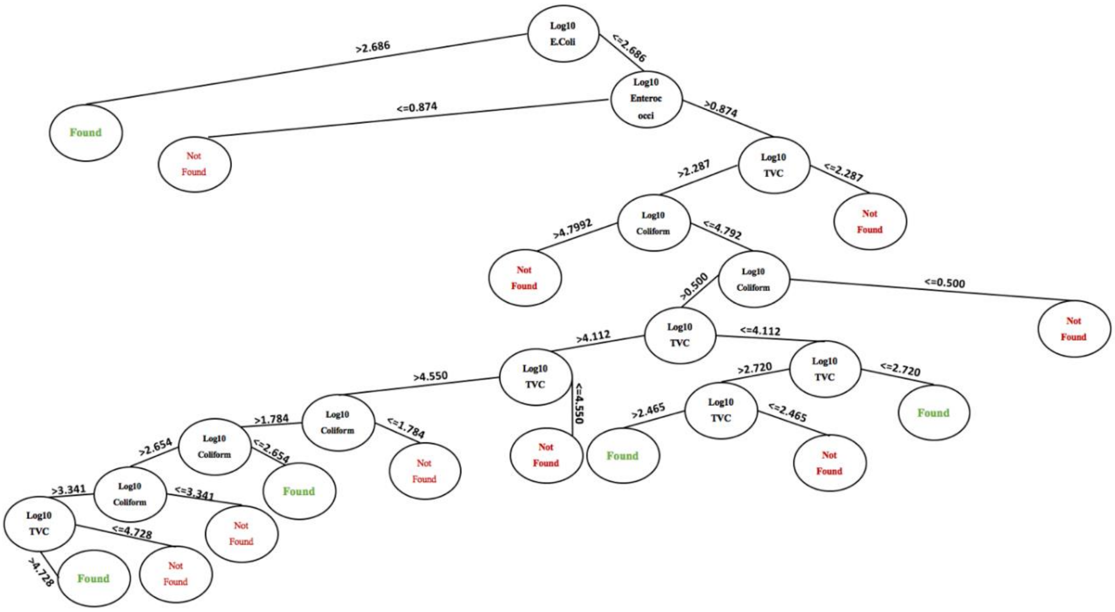


Figure 14. The second generated tree in cooked chicken dataset

For the second generated tree in cooked chicken dataset, it composes all features having *E.coli* as the root node.

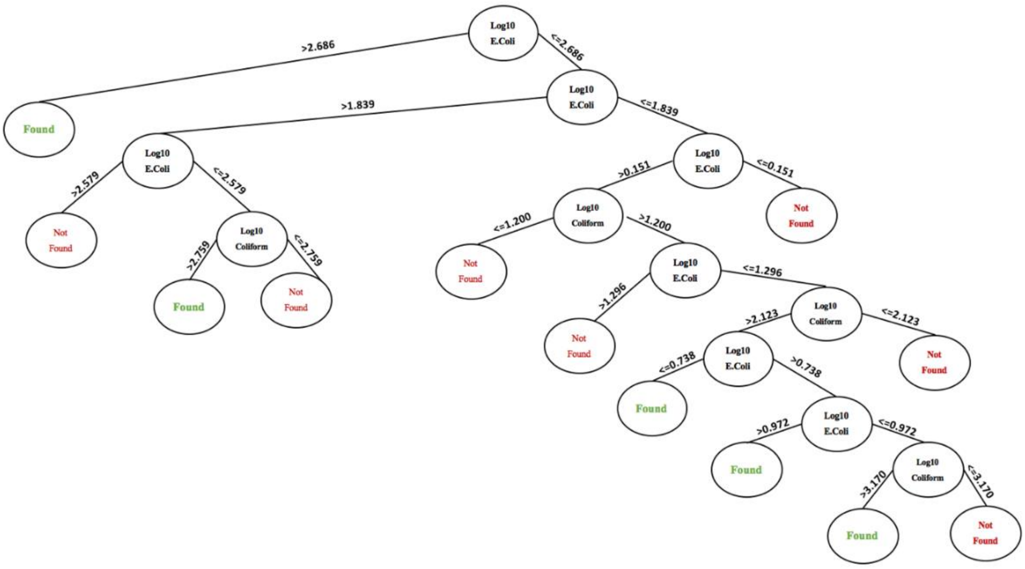


Figure 15. The third generated tree in cooked chicken dataset

For the third generated tree in cooked chicken dataset, there are 2 features, namely, *E.coli* and *Coliform.* where, *E.coli* is used as the root node.

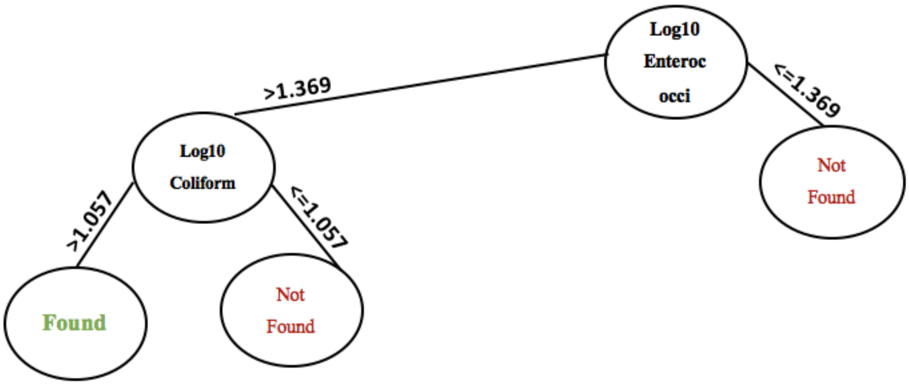


Figure 16.The fourth generated tree in cooked chicken dataset

For the fourth generated tree in cooked chicken dataset, there are 2 features, namely, *Enterococci* and *Coliform.*

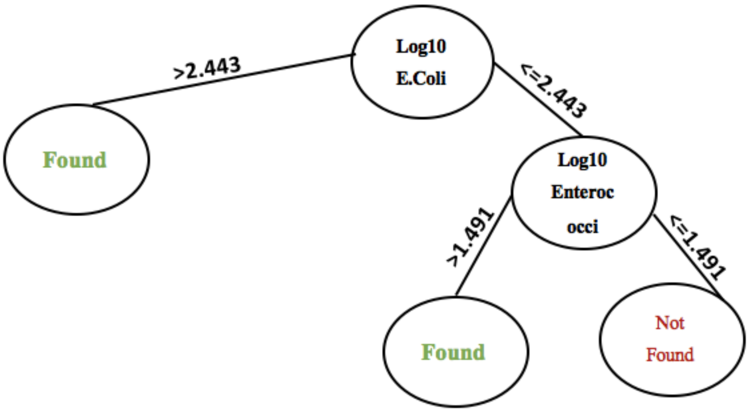


Figure 17. The fifth generated tree in cooked chicken dataset

For the fifth generated tree in cooked chicken dataset, there are two features, namely, *E.coli* and *Enterococci*, where *E.coli* is the root node.

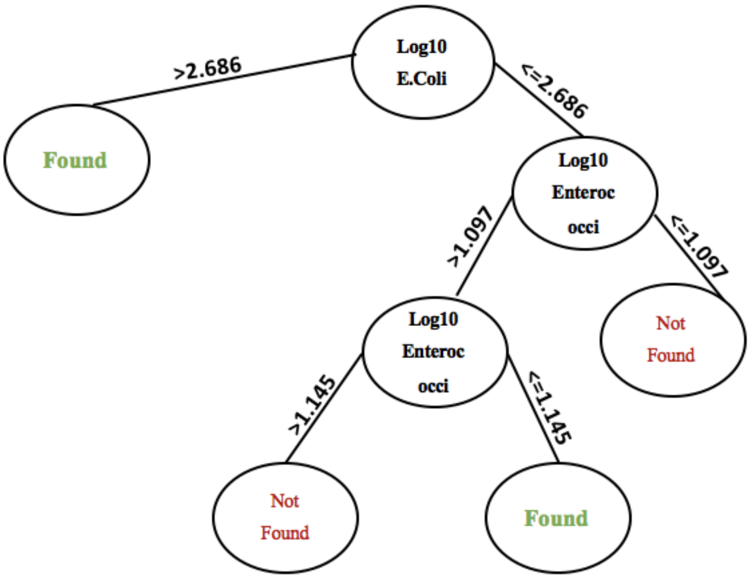


Figure 18. The sixth generated tree in cooked chicken dataset

For the sixth generated tree in cooked chicken dataset, there are two features, namely, are *E.coli* and *Enterococci*.

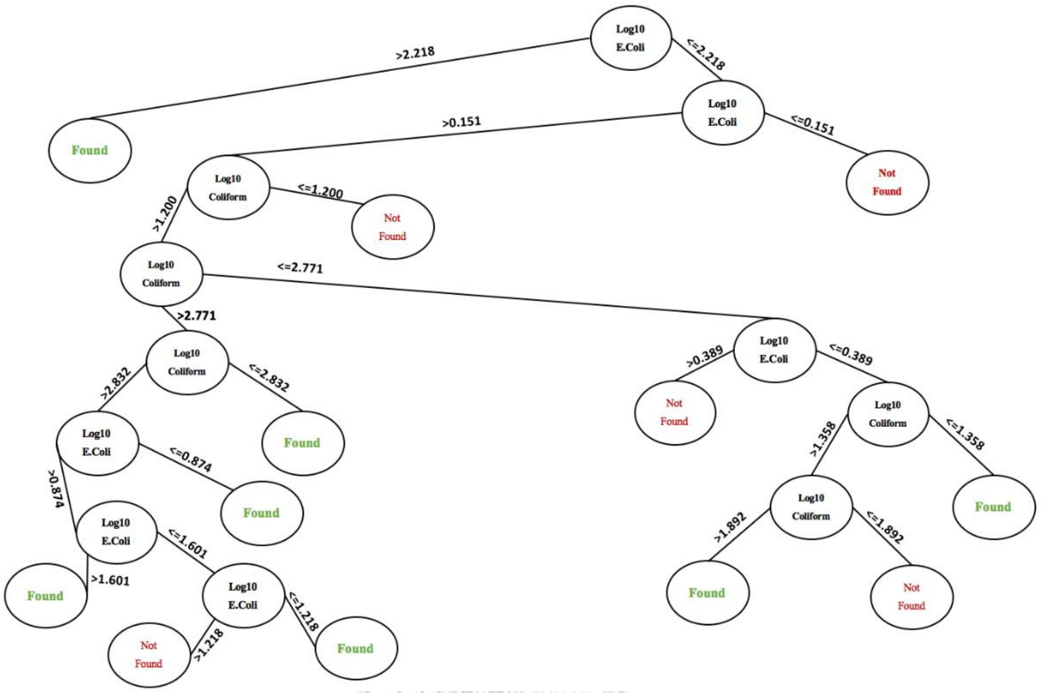


Figure 19. The seventh generated tree in cooked chicken dataset

For the seventh generated tree in cooked chicken dataset, there are two features, namely, *E.coli* and *Coliform,* where *E.coli* is the outstanding feature.

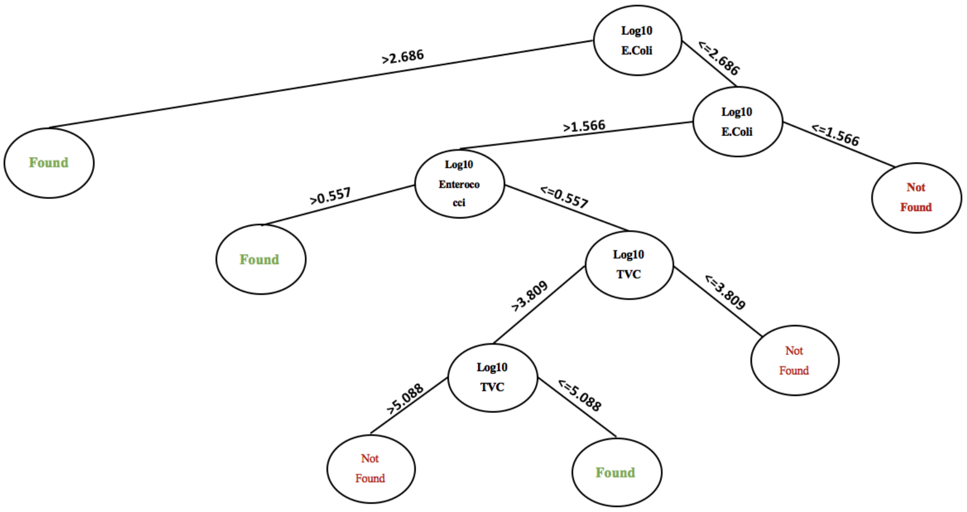


Figure 20. The eighth generated tree in cooked chicken dataset

For the eighth generated tree in cooked chicken dataset, there are three features, namely, *E.coli, Enterococci and TVC,* where *E.coli is the root node.*

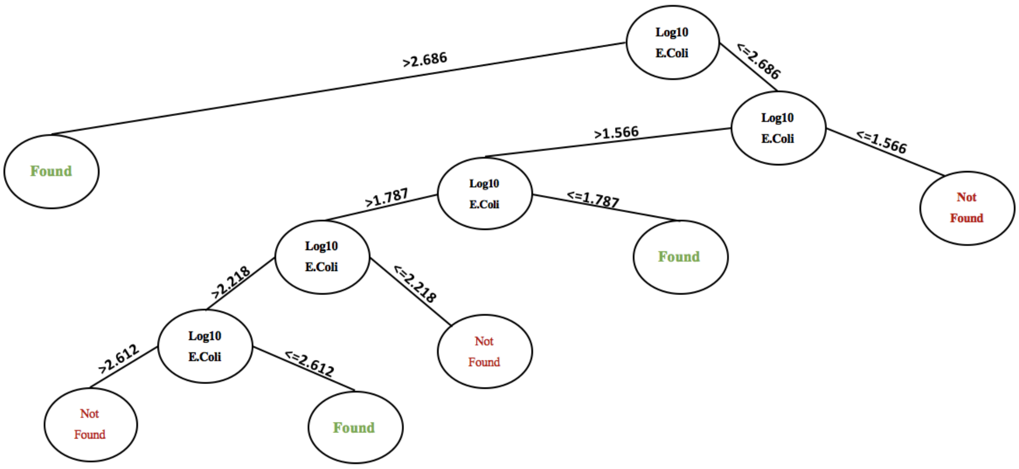


Figure 21. The ninth generated tree in cooked chicken dataset

For the ninth generated tree in cooked chicken dataset, *E.coli* is the only feature.

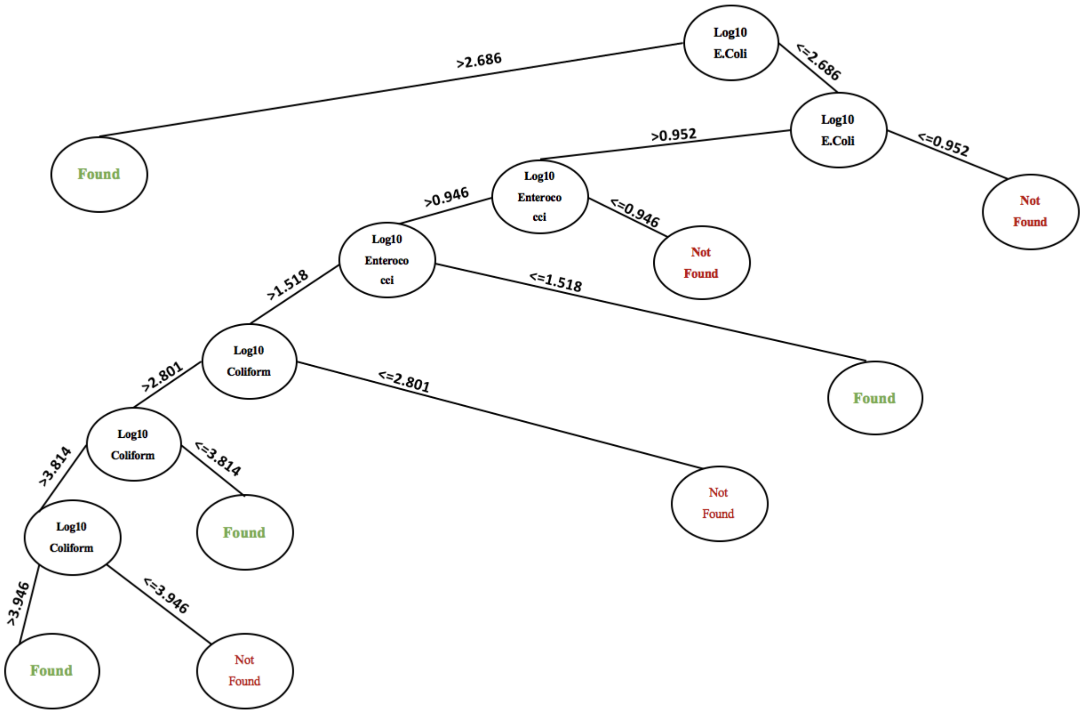


Figure 22. The tenth generated tree in cooked chicken dataset

For the tenth generated tree in cooked chicken dataset, there are three features, namely, *E.coli, Enterococci* and *Coliform, respectively.*

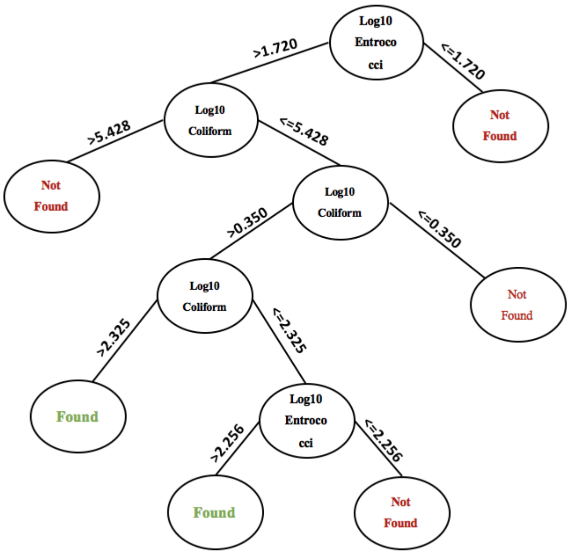


Figure 23. The eleventh generated tree in cooked chicken dataset

For the eleventh generated tree in cooked chicken dataset, it composes of 2 features, namely, *Enterococci* and *Coliform*.

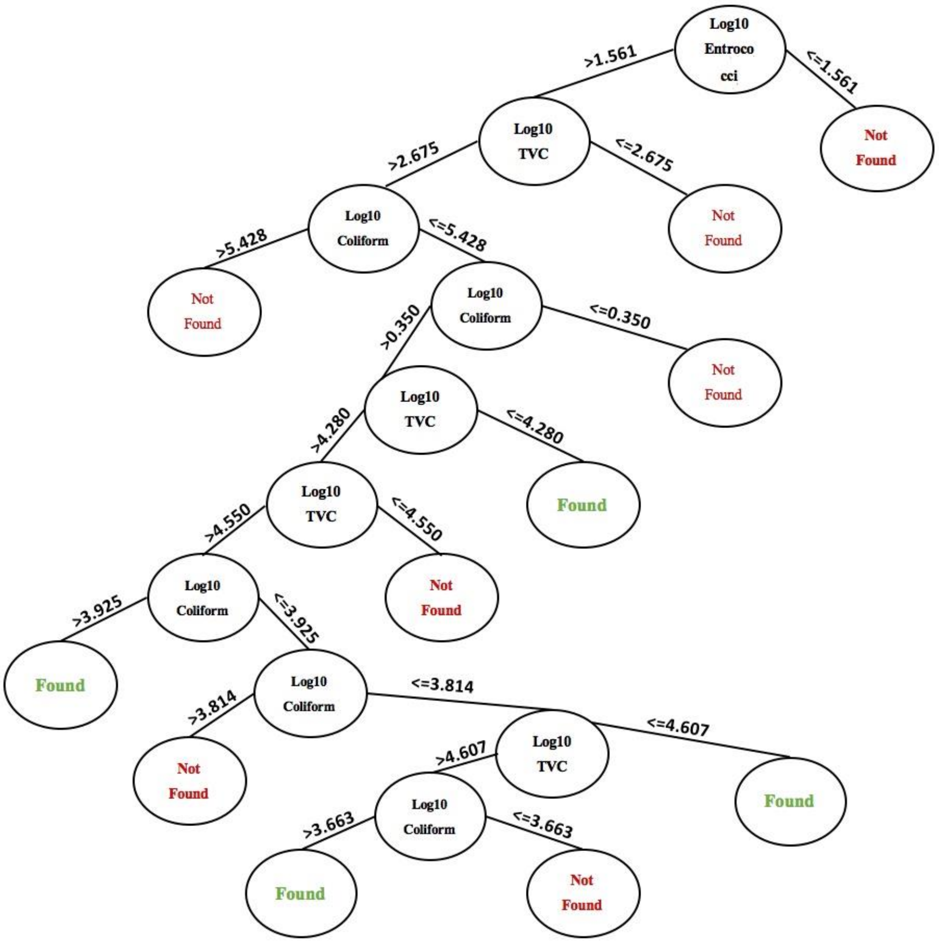


Figure 24. The twelfth generated tree in cooked chicken dataset

For the twelfth generated tree in cooked chicken dataset, there are three features, namely, *Enterococci*, *Coliform,* and *TVC*, where *Enterococci* is the root node.

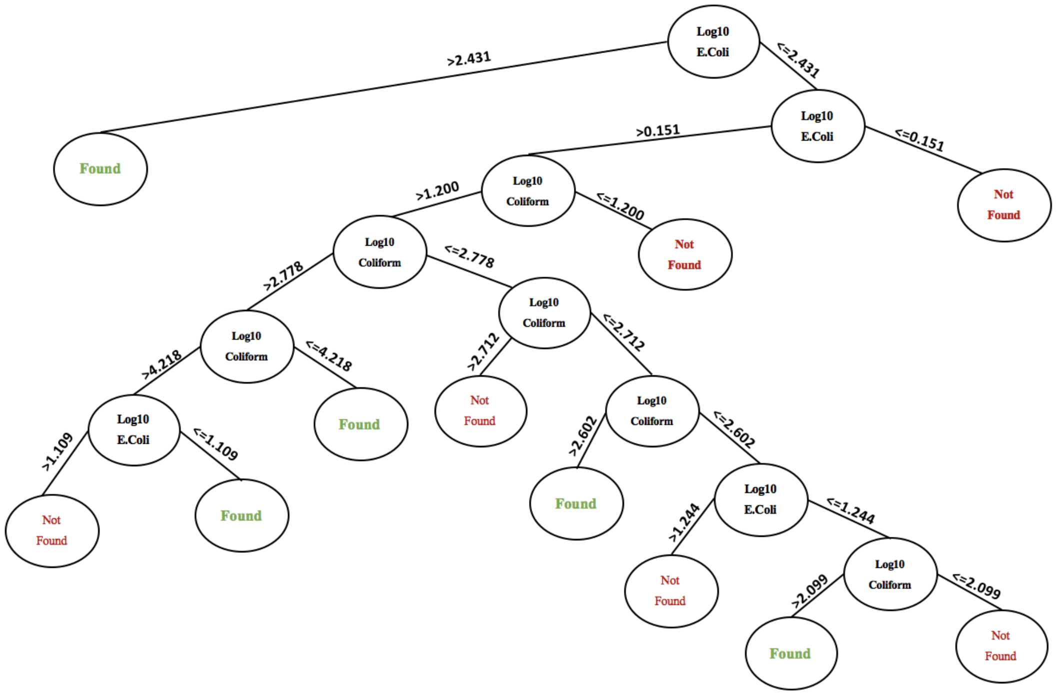


Figure 25. The thirteenth generated tree in cooked chicken dataset

For the thirteenth generated tree in cooked chicken dataset, there are two features, namely, *E.coli* and *Coliform*, where, *E.coli* is the root node.

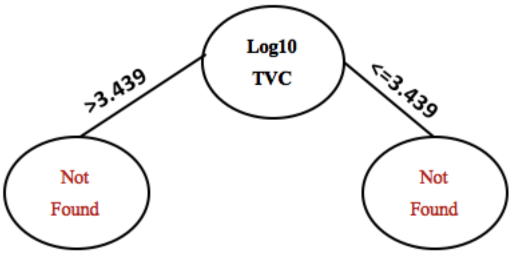


Figure 26. The fourteenth generated tree in cooked chicken dataset

For the fourteenth generated tree in cooked chicken dataset, there is only one feature which is *TVC*. The result of both sides are negative classes.

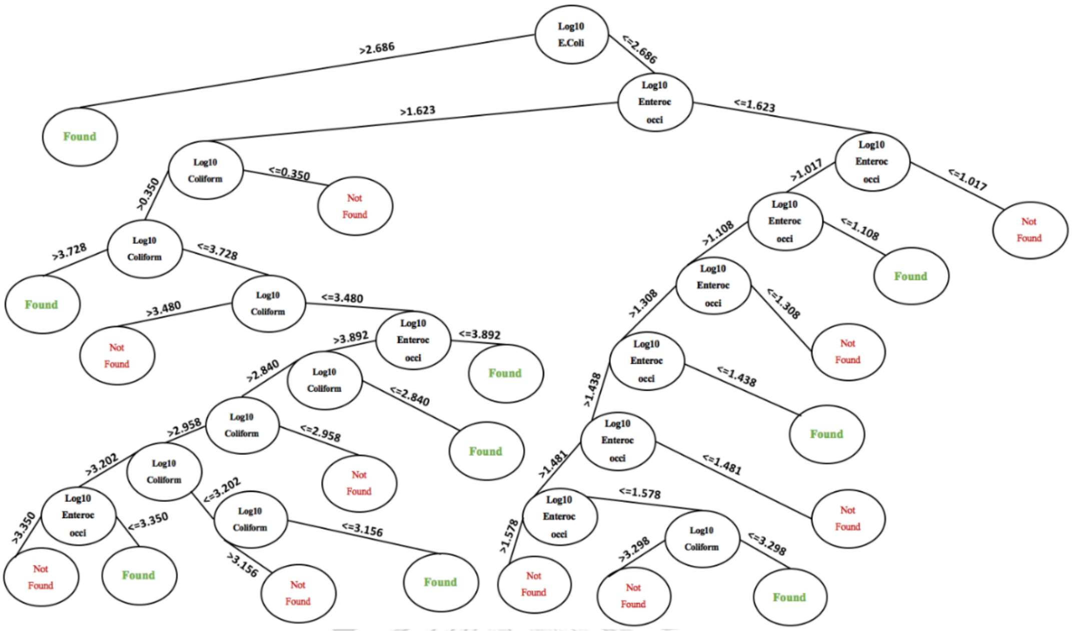


Figure 27. The fifteenth generated tree in cooked chicken dataset

For the fifteenth generated tree in cooked chicken dataset, it represents a huge random tree and *E.coli*, *Enterococci,* and *Coliform* are features.

3.2.3 Finding relationship among trees

After the trees are generated, the sets of trees that has similarity of structure are found in 7 random trees namely, tree 1, 2, 7, 8, 10, 12, and 15. Next, the condition paths which obtain many positive classes are listed and merged the paths that have the same condition consecutively. The selected condition paths are shown in Figure 28. to Figure 31.

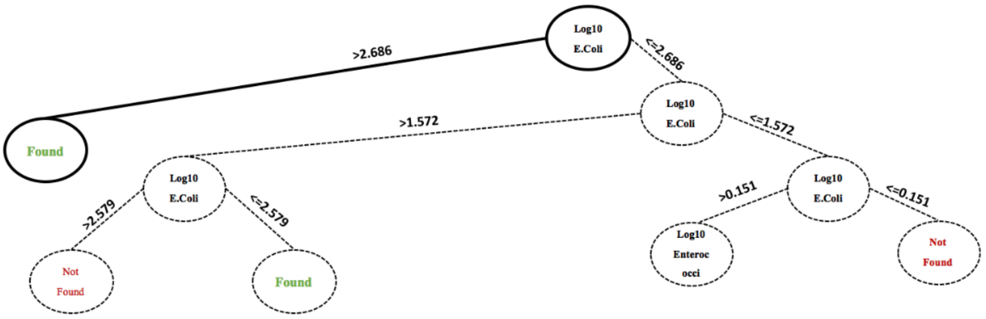


Figure 28. The first selected condition path in cooked chicken dataset

For the first selected condition path in cooked chicken dataset*,* *E.coli* is greater than 2.686 to *reach the positive class.*

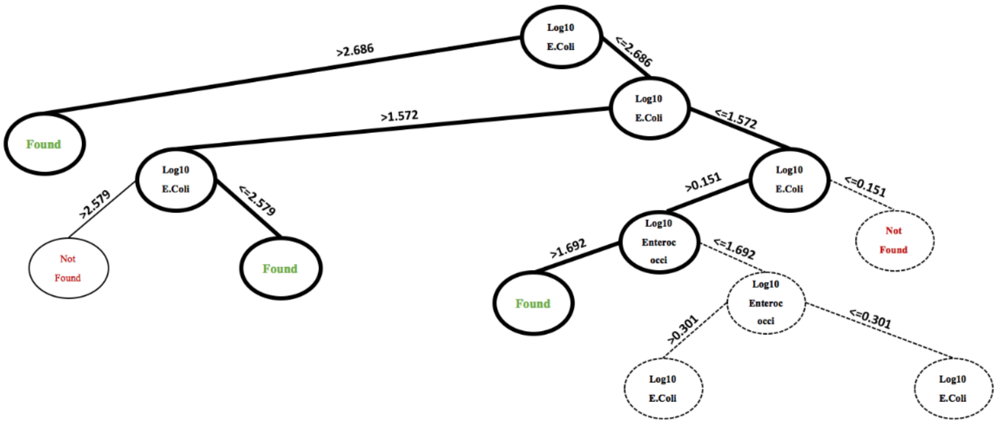


Figure 29. The second selected condition path in cooked chicken dataset

For the second selected condition path in cooked chicken dataset, it starts the value of *E.coli* over *0.151,* followed by *Enterococci* over 1.682, constituting the positive class.

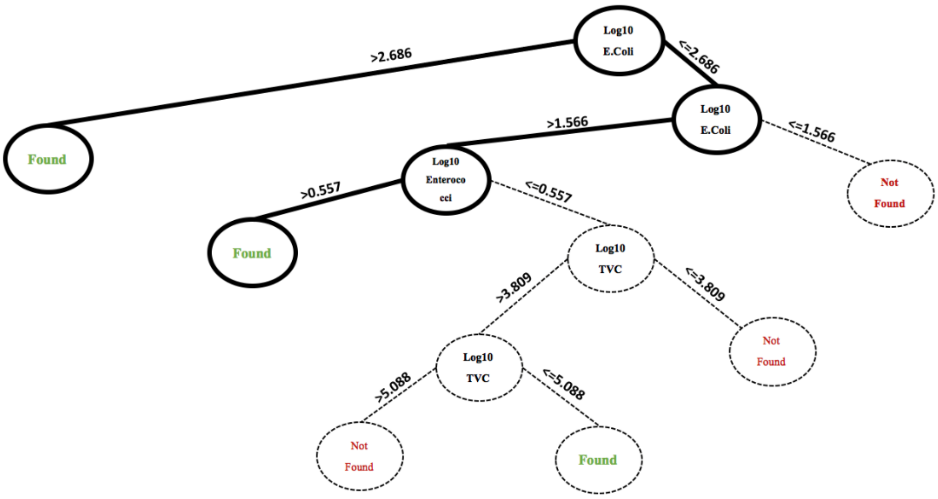


Figure 30. The third selected condition path in cooked chicken dataset

For the third condition path in cooked chicken dataset, it starts from the value of *E.coli* over 1.566, then *Enterococci* over 0.557 to reach the positive class.

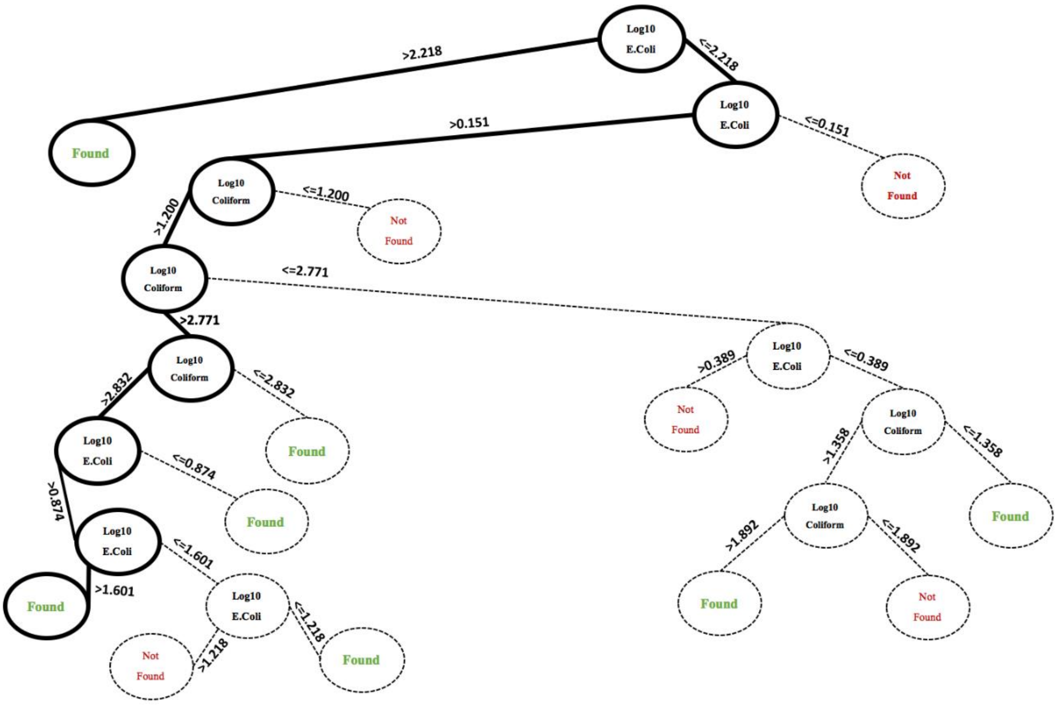


Figure 31. The fourth selected condition path in cooked chicken dataset

For the fourth selected condition path in cooked chicken dataset, it starts from the value of *E.coli* over 0.151, then *Coliform* over 2.832.

The above condition paths show the important features and conditions that are key factors to cause *Listeria* contamination. The results obtained from random forest method for cooked chicken dataset compose of four conditions as shown in Table 1.

Table 1. The selected condition paths from random forest

Model	Quantity of Bacterium			
	E.coli	*Enterococci*	*Coliform*	*TVC*
1st condition	>0.752	>1.167	-	-
2nd condition	>0.151	>1.692	-	-
3rd condition	>1.566	>0.557	-	-
4th condition	>0.151	-	>2.832	-

The first condition represents *E.coli* greater than 0.752 and *Enterococci* greater than 1.167. The second condition represents *E.coli* greater than 0.151 and *Enterococci* greater than 1.692. The third condition represents *E.coli* greater than 1.566 and *Enterococci* greater than 0.557. The fourth condition represents *E.coli* greater than 0.151 and *Coliform* greater than 2.832. All these conditions represent the units of bacteria in log 10 form. As the result condition, *E.coli* is the outstanding factor to cause *Listeria* contamination.

These resulting conditions are selected because the number of positive classes dominate the negative ones as shown in Table 2.

Table 2. The number of results for positive class and negative class

Model	Result	
	Positive Class	*Negative Class*
1st condition	20	4
2nd condition	18	1
3rd condition	16	2
4th condition	21	6

3.3 Linear Regression

3.3.1 Applied linear regression

Using the features and condition paths of *Listeria* contamination obtained from random forest, Linear regression is applied to refine the results. First of all, the same dataset used in random forest is applied with linear regression for the numerical prediction. The equation becomes

$$Y = (0.034)x_1 + (0.193)x_2 + (0.105)x_3 + 0.035 \qquad (1)$$

where x_1 denotes log10(*Coliform*), x_2 denotes log10(*E.coli*), and x_3 denotes log10(*Enterococci*).

3.3.2 Overlapping between random forest result and linear regression result

This step is to ensure the results from random forest, the record overlapping between random forest and linear regression is used. Record overlapping will be used to compare to every single conditions in random forest manually. The set of record from the equation of linear regression is listed. The path that gets the highest best match records will be selected by setting up the threshold of linear regression. The threshold is set at 0.737 to permit the best matching between random forest and linear regression that will yield appropriate number of positive classes.

After record the overlapping between random forest and linear regression, the number of positive class and negative class change as shown in Table 3.

Table 3. The number of result for positive class and negative class after overlapped the result

Model	Random Forest		Linear Regression	
	Positive Class	*Negative Class*	*Positive Class*	*Negative Class*
1st condition	20	4	11	1
2nd condition	18	1	10	0
3rd condition	16	2	11	1
4th condition	21	6	11	1

In first condition, the result of positive class is 11 and negative class is 1. The second and the remaining value are (10,0), (11,1), and (11,1) respectively. The rational results of the first, third, and fourth conditions is that there are the same number of positive classes and negative classes. Apparently, the first condition and the third condition have the same features and the number of records. Thus, both conditions are merged to integrate these two conditions into one condition. This condition denotes *E.coli* to be greater than 1.672 and *Enterococci* greater than 1.362. There are two relationships that cause *Listeria* contamination i.e., *E.coli*, *Enterococci* and *E.coli,* and *Coliform*. Thus, the new condition and the fourth condition are retained as shown in Table 4.

Table 4. The selected conditions after refinement

Model	Quantity of Bacterium			
	E.coli	*Enterococci*	*Coliform*	*TVC*
New condition	>1.672	>1.362	-	-
4th condition	>0.151	-	>2.832	-

In summary, since the coefficient form of linear regression conforms to the significant features from random forest result, both results indicate that *E.coli* is the most significant factor to cause *Listeria* contamination. Obviously, *E.coli* is the root of all four conditions and gets the highest weight in coefficient of Linear regression equation. Then, *Enterococci* is the next child node in random forest tree result and gets the second highest weight in coefficient of Linear regression equation. *Coliform* is one of the variants in the linear regression equation despite small weight in coefficient. In contrast, there is no *TVC* in the weight of coefficient and tree construction.

3.4 Naïve Bayes

After getting two conditions from linear regression, this step reassures the proposed model. First step is the data preparation. The data consist of training set and test set. There are 12 records from linear regression consisting of 11 positive classes and 1 negative class. They are selected to be the test set. For the training set, the same dataset from random forest and linear regression except the above twelve test sets is used. Then, Naïve Bayes is applied on the training set and test set. The result of Naïve Bayes is the same as that of linear regression.

After getting the validation the result by Naïve Bayes, the two conditions become specific number of bacteria units on cooked chicken dataset, so those retained conditions are generalized to be the proposed model. Normalization method is used as shown in equation (2).

$$\boxed{\textit{Normalized valued} \quad = \quad \frac{\textit{Exact Value} - \textit{Minimum Value}}{\textit{Maximum Value} - \textit{Minimum Value}}} \qquad (2)$$

Table 5 shows the proposed model after normalization. The first condition denotes *E.coli* to be greater than 0.479 and *Enterococci* greater than 0.5678. The second condition denotes *E.coli* to be greater than 0.0429 and *Coliform* greater than 0.663.

Table 5. The proposed model condition after generalization

Model	Quantity of Bacterium			
	E.coli	*Enterococci*	*Coliform*	*TVC*
1st condition	>0.479	>.5879	-	-
2nd condition	>0.0429	-	>0.663	-

Chapter 4.
Evaluation Methodology

The proposed model is evaluated by another dataset from the food expert which is cooked rice dataset. The way to evaluate is the result comparison between the proposed model and single classification. The cooked rice dataset has three features, namely, *TVC*, *E.coli,* and *Coliform*. Obviously, there is no *Enterococci* in the features that is the reason why the only second condition is used for model evaluation. The second condition in the proposed model shows that *E.coli* is greater than 0.0429 and *Coliform* is greater than 0.663. To apply the proposed model with cooked rice dataset, value range mapping between cooked rice dataset and proposed model is performed. Then, *E.coli* adjusted to be greater than 0.730 and *Coliform* greater than 2.787 in cooked rice dataset. As a consequence, the result shows 6 positive classes and 14 negative classes after applying the proposed model with cooked rice dataset as shown in Table 6.

Model	**Result**	
	Positive Class	***Negative Class***
Proposed model	6	14

Table 6. The number of result for positive class and negative class after applied the proposed model in cooked rice dataset

The proposed model utilized classification refinement technique by using single classification in single random forest and single linear regression.

4.1 Comparison between proposed model and single random forest

For random forest, the initial settings select features as gain ratio, max depth as 20 and tree as 15. Only 14 random trees are generated as shown in Figure 32. to Figure 45. The other tree cannot represent the result that could be affected by the number of records in dataset.

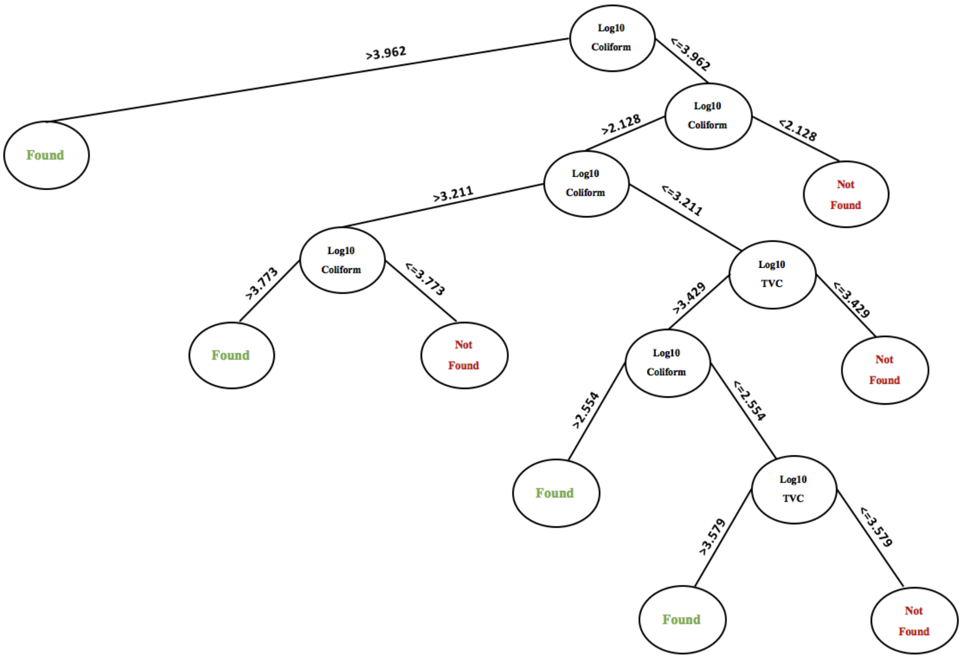


Figure 32. The first generated tree in cooked rice dataset

For the first generated tree in cooked rice dataset, there are two features, namely, *Coliform* and *TVC*, where *Coliform* is the root node.

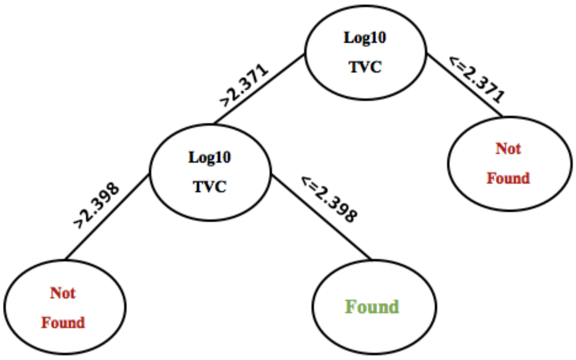


Figure 33. The second generated tree in cooked rice dataset

For the second generated tree in cooked rice dataset, *TVC* is the only one feature.

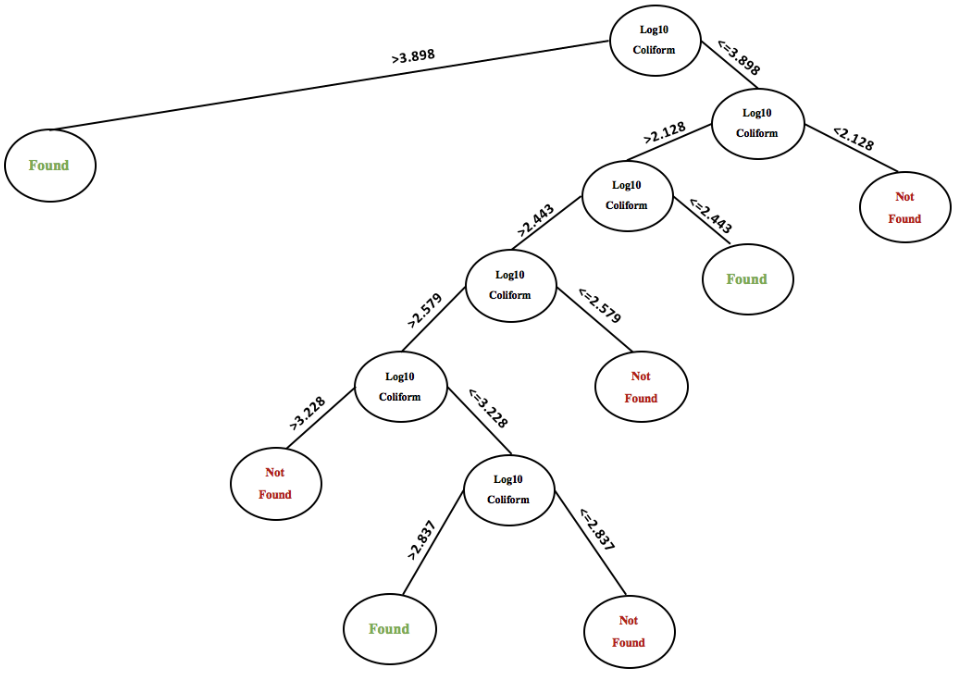


Figure 34. The third generated tree in cooked rice dataset

For the third generated tree in cooked rice dataset, *Coliform* is the only one feature in this tree.

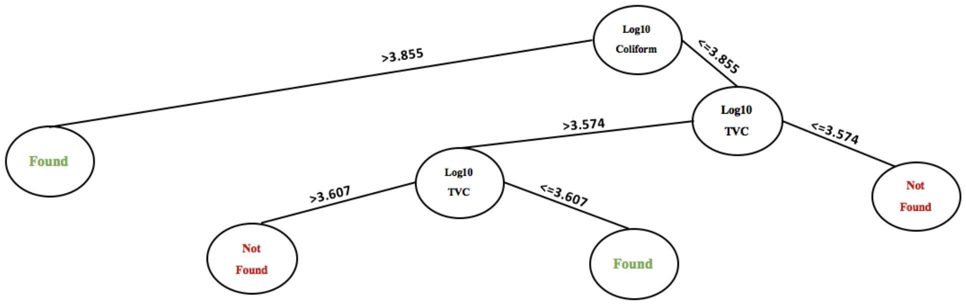


Figure 35. The fourth generated tree in cooked rice dataset

For the fourth generated tree in cooked rice dataset, there are two features, namely, *Coliform* and *TVC*, where *Coliform* is the root node.

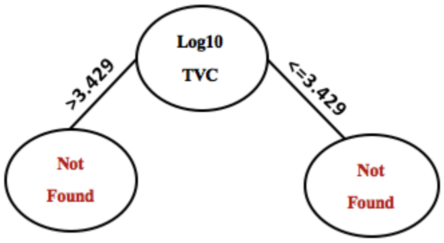


Figure 36. The fifth generated tree in cooked rice dataset

For the fifth generated tree in cooked rice dataset, *TVC* is the only one features in this tree and the class result for both sides is negative class.

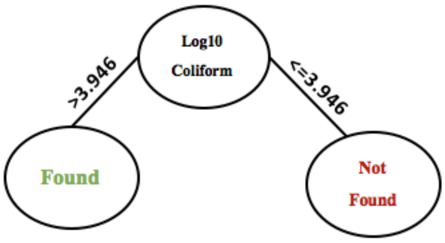


Figure 37. The sixth generated tree in cooked rice dataset

For the sixth generated tree in cooked rice dataset, *Coliform* is the only one feature. *Coliform* is greater than 3.946 to get the positive result.

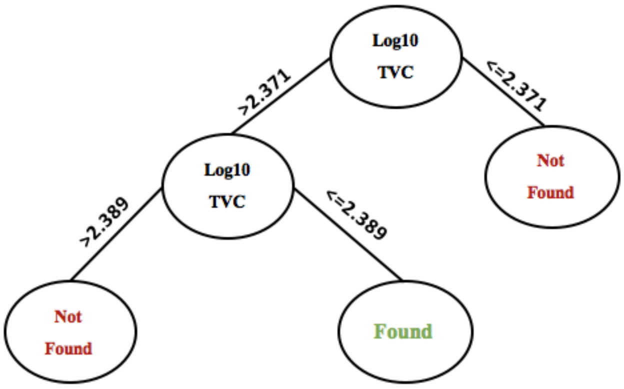


Figure 38. The seventh generated tree in cooked rice dataset

For the seventh generated tree in cooked rice dataset, there are only *TVC* that is the feature.

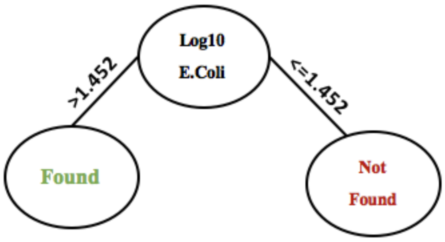


Figure 39. The eighth generated tree in cooked rice dataset

For the eighth generated tree in cooked rice dataset, *E.coli* is the only one feature and is greater than 1.452 to reach the positive class.

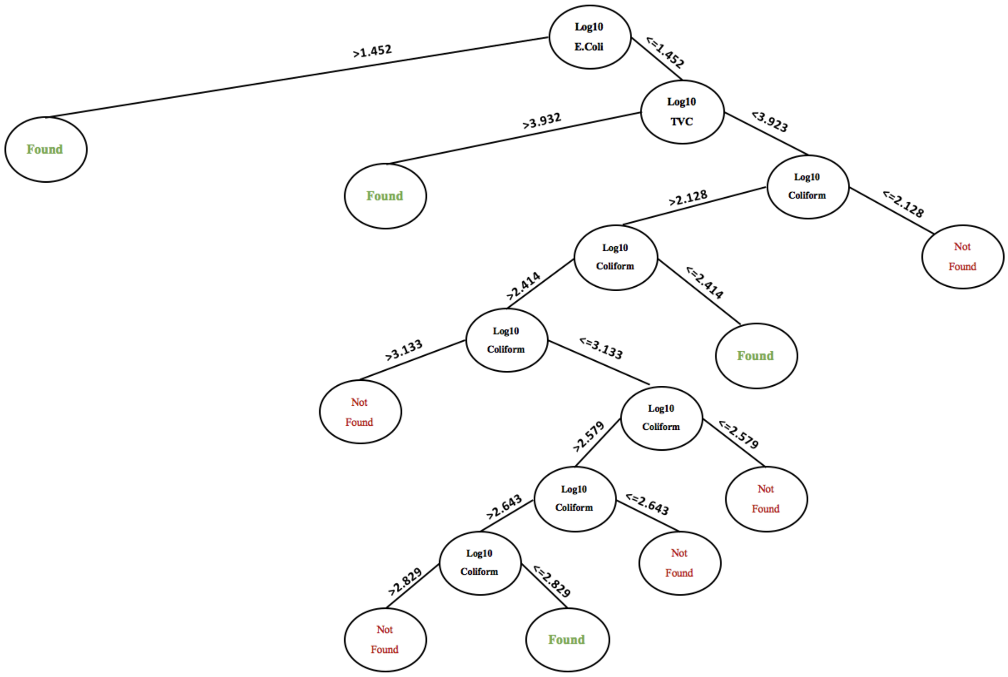


Figure 40. The ninth generated tree in cooked rice dataset

Foe the ninth generated tree in cooked rice dataset, there are three features, namely, *E.coli*, *TVC,* and *Coliform*.

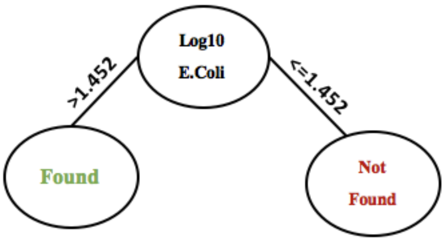


Figure 41. The tenth generated tree in cooked rice dataset

For the tenth generated tree in cooked rice dataset, *E.coli* is the only one feature and greater than 1.452 to get the positive class.

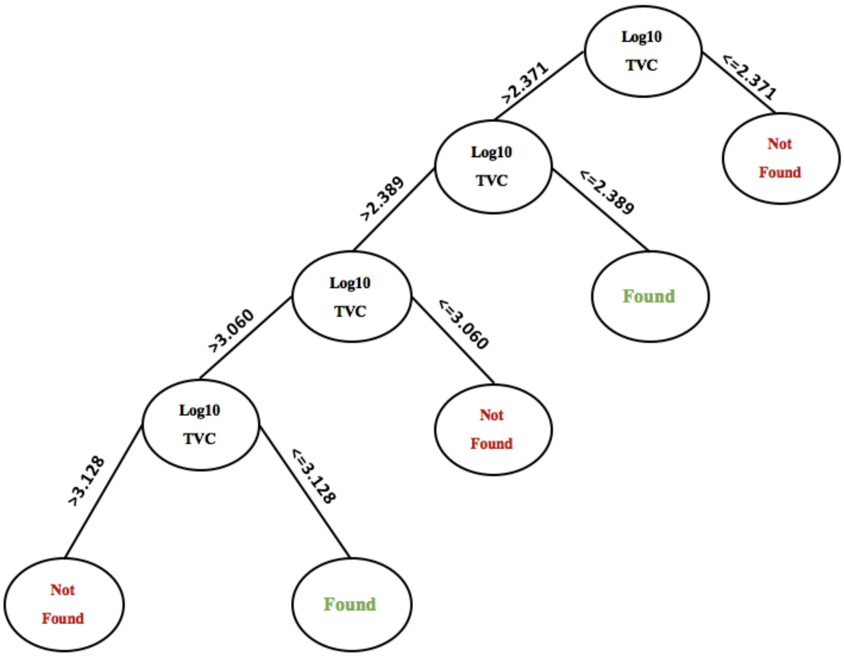


Figure 42. The eleventh generated tree in cooked rice dataset

For the eleventh generated tree in cooked rice dataset, *TVC* is only one feature in this tree.

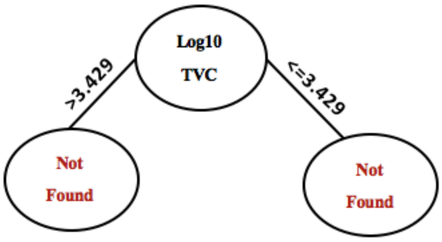


Figure 43. The twelfth generated tree in cooked rice dataset

For the twelfth generated tree in cooked rice dataset, *TVC* is the only one features in this tree and the class result for both sides is negative class.

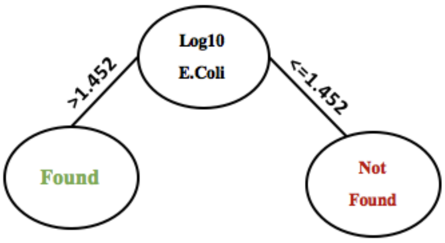


Figure 44. The thirteenth generated tree in cooked rice dataset

For the thirteenth generated tree in cooked rice dataset, *E.coli* is the only one feature and greater than 1.452 to get the positive class.

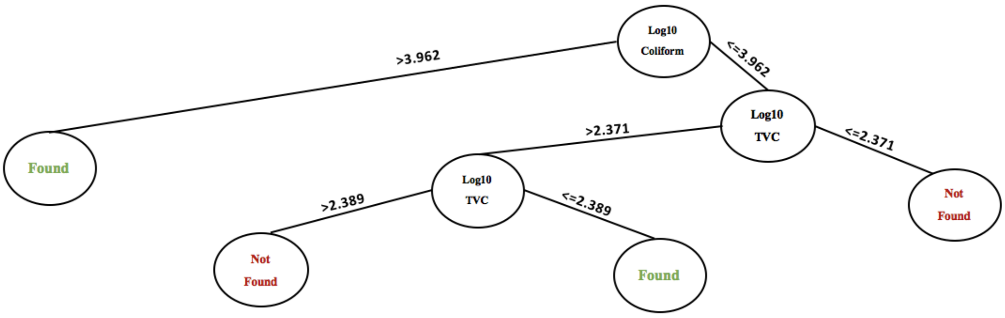


Figure 45. The fourteenth generated tree in cooked rice dataset

For the fourteenth generated tree in cooked rice dataset, there are two features, namely, *Coliform* and *TVC*.

The result tree varies less in condition path which may be caused by the limitation of dataset having less number of records. After completing generating trees, the most positive classes are selected, the result represents two condition paths. For the first condition path, *E.coli* is only factor to cause *Listeria* contamination. There are 3 positive classes and 1 negative class in the result. For the second condition path, *Coliform* is the only one variant that causes *Listeria* contamination. There are 3 positive classes and 2 negative classes in result. It means that the proposed method returns better number of positive classes than single random forest as shown in Table 7.

Table 7. The comparison number of result between proposed model and single random forest

Model	Result	
	Positive Class	*Negative Class*
Proposed model	6	14
Single Random forest	3	2

4.2 Comparison between proposed model and single linear regression

Single linear regression is selected to evaluate the proposed model. There are two steps, namely, data preparation for 3-fold cross validation and applying to each training set. The dataset and test set are uniquely separated into 3 parts for 3-fold cross validation. The cooked rice dataset has 159 records consisting of 17 positive classes and 143 negative classes. To set the similar quantity of positive classes, the positive classes of cooked rice dataset is divided into 6 positive classes per fold. The proportion of positive classes and negative classes are 6:47,6:47, and 5:48, respectively. All those proportions are separated to be the test set. The result is about 1 or 2 positive classes as shown in Table 8.

Table 8. The number of result in each fold for linear regression

Model	Result	
	Positive Class	*Negative Class*
1st Fold	1	50
2nd Fold	1	50
3rd Fold	2	51

It means the proposed model gets more positive classes at 6 positive classes as shown in Table 9.

Table 9. Comparison of number result among proposed model, single random forest and single linear regression

Model	Result	
	Positive Class	*Negative Class*
Proposed model	6	14
Single Random forest	3	2
Single Linear regresion	2	51

Table 10 shows the records result class from the test set which composes of 20 records. Those records include 6 positive classes and 14 negative classes. The rest of cooked rice dataset is used to be the training set. The Naïve Bayes result shows the same outcome from the proposed method.

Table 10. The number of result after evaluated the proposed model in cooked rice dataset

Model	**Result**	
	Positive Class	***Negative Class***
Proposed model	6	14
Naïve Bayes	6	14

To sum up, based on the compared results, the proposed method which is the classification refinement technique including random forest and linear regression, has the capability to find the actual related bacteria indicators and conditions of bacteria unit to lead to *Listeria* contamination. The results indicate that *E.coli* is the important factor of causing *Listeria* contamination for both cooked chicken dataset and cooked rice dataset. Additionally, the combinations of bacteria likes *E.coli* with *Enterococci* and *E.coli* with *Coliform*, are also able to cause *Listeria* contamination.

Chapter 5.
DISCUSSION AND CONCLUSION

5.1 Discussion

According to the model and the evaluation results, it can be seen that the merging classification method is able to refine the actual related bacteria which causes *Listeria* contamination. The based component feature of random forest method supports the proposed model to find the important factors in each tree path. The main path can represent the condition of the bacteria unit to get the resulting class. Thus, random forest results compose of the important factors and the condition path. Linear regression is applied to refine the random forest results. As the based concept of linear regression, the equation of linear regression is obtained. The equation composes of the coefficient of each variant and the variants. Thus, the coefficients of variant represent the essential weight of the particular variant.

Based on the concept of both random forest and linear regression, result of the proposed model obtain significant factors and significant condition tree paths from random forest, while eliminating useless conditions by linear regression. Thus, after refinement, the important condition paths are merged and adjusted the value. To validate the proposed model, the conclusive condition paths are evaluated by result comparison. The results consist of the proposed model, single random forest, and single linear regression. The first comparison uses single random forest that yields too broad range of random forest. Thus, the proportions of positive class and negative class are different. If the number of positive class is a small number. It may affect the average result.

The next comparison uses the single linear regression that yields good results to indicate the relating factor to *Listeria* contamination. However, it is unable to describe the quantity of the bacteria to create *Listeria* contamination as the condition path from random forest. Obviously, classification refinement model is able to support finding the related bacteria factor which lead to *Listeria* contamination.

5.2 Conclusion

The proposed classification refinement technique combines two classification methods which are random forest and linear regression. The consecutive embedded classification is able to refine the actual significant results and the conditions of the bacteria quantity that cause *Listeria* contamination. The results of the proposed model yield the superior outcome than using single random forest or single linear regression, that is, proper number of positive classes.

REFERENCES

[1] T. B. P. E. Association. (2017). *Thai chicken meat export[online]*. Available: http://www.thaipoultry.org/welcome.php?page=ExportStatic.

[2] D. o. L. Development. (2010). *Microbiological criteria of livestock products for export [online].* Available: http://www.dld.go.th/certify/th/index.php?option1/4com_content&view1/4article&id1/4609:law&catid1/4118:docuqmentary

[3] Y. Ma, Y. Hou, Y. Liu, and Y. Xue, "Research of food safety risk assessment methods based on big data," in *Big Data Analysis (ICBDA), 2016 IEEE International Conference on*, 2016, pp. 1-5: IEEE.

[4] Y. Ma, J. Liu, Y. Lin, X. Cai, and Z. Liu, "Research and exploration of the Key Elements of Food Safety Data Analysis System based on the food safety traceability system," in *Computer Science & Education (ICCSE), 2014 9th International Conference on*, 2014, pp. 601-605: IEEE.

[5] R. Zhang, S. Zhao, Z. Jin, N. Yang, and H. Kang, "Application of SVM in the food bacteria image recognition and count," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, 2010, vol. 4, pp. 1819-1823: IEEE.

[6] C. Zheng, J. Liu, and G. Qiu, "Tuberculosis bacteria detection based on Random Forest using fluorescent images," in *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on*, 2016, pp. 553-558: IEEE.

APPENDIX

VITA

Name: Napas Jeamchotpatanakul

Affiliation: Advanced Virtual and Intelligent Computing (AVIC) Center, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University.

Country: Thailand

Biography: Miss Napas Jeamchotpatanakul was born on July 6, 1992, in Bangkok province, Thailand. She received a Bachelor's degree in Computer Science and Information Technology from Thammasat University. Now she is a Master's degree student in Computer Science and Information Technology, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University.