ESTIMATING PHENYLALANINE OF COMMERCIAL FOODS : A COMPARISON BETWEEN A MATHEMATICAL APPROACH AND A MACHINE LEARNING APPROACH

A Thesis

Submitted to the Faculty

of

Purdue University

by

Amruthavarshini Talikoti

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Electrical and Computer Engineering

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF THESIS APPROVAL

Dr. Mireille Boutin, Chair
School of Electrical and Computer Engineering
Dr. Edward J. Delp
School of Electrical and Computer Engineering
Dr. Fengqing M. Zhu
School of Electrical and Computer Engineering

Approved by:
Dr. Pedro Irazoqui
Head of the School Graduate Program

To my parents, Appa and Amma

&

my parent at Purdue, Professor Mimi.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude and indebtedness to Professor Mireille Boutin for her constant support and invaluable guidance in the past two years. I would like to specially acknowledge her attention to detail, encouraging mentoring and an understanding attitude. With a patient and kind demeanor, she played a major role in nurturing my growth as a learner, presenter and researcher. I sincerely believe that my greatly positive experience of Masters thesis research is credited to her. Working with Prof. Boutin has reinforced my strong desire to pursue a PhD in the near future; for this and many more experiences, I am highly indebted to her.

I would like to thank Jieun Kim based on whose work my research was built. Her invaluable suggestions have given us key directions to progress. I would like to acknowledge her well-written documents and codes that helped me gain a good understanding of the project. I would also like to thank her for actively responding to the project even after her stay at Purdue.

I would like to thank my committee members Professor Edward J. Delp and Professor Fengqing M. Zhu for their invaluable insights and for having made the experience smooth.

I would like to thank my family and friends for their encouraging love and support that has successfully helped me reach this milestone.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Talikoti, Amruthavarshini M.S., Purdue University, May 2019. Estimating Phenylalanine of Commercial Foods : A Comparison Between a Mathematical Approach and a Machine Learning Approach. Major Professor: Mireille Boutin.

Phenylketonuria (PKU) is an inherited metabolic disorder affecting 1 in every 10,000 to 15,000 newborns in the United States every year. Caused by a genetic mutation, PKU results in an excessive build up of the amino acid Phenylalanine (Phe) in the body leading to symptoms including but not limited to intellectual disability, hyperactivity, psychiatric disorders and seizures. Most PKU patients must follow a strict diet limited in Phe. The aim of this research study is to formulate, implement and compare techniques for Phe estimation in commercial foods using the information on the food label (Nutritional Fact Label and ordered ingredient list). Ideally, the techniques should be both accurate and amenable to a user friendly implementation as a Phe calculator that would aid PKU patients monitor their dietary Phe intake.

The first approach to solve the above problem is a mathematical one that comprises three steps. The three steps were separately proposed as methods by Jieun Kim in her dissertation. It was assumed that the third method, which is more computationally expensive, was the most accurate one. However, by performing the three methods subsequently in three different steps and combining the results, we actually obtained better results than by merely using the third method.

The first step makes use of the protein content in the foods and Phe:protein multipliers. The second step enumerates all the ingredients in the food and uses the minimum and maximum Phe:protein multipliers of the ingredients along with the protein content. The third step lists the ingredients in decreasing order of their weights, which gives rise to inequality constraints. These constraints hold assum-

ing that there is no loss in the preparation process. The inequality constraints are optimized numerically in two phases. The first involves nutrient content estimation by approximating the ingredient amounts. The second phase is a refinement of the above estimates using the Simplex algorithm. The final Phe range is obtained by performing an interval intersection of the results of the three steps. We implemented all three steps as web applications. Our proposed three-step method yields a high accuracy of Phe estimation (error $\leq \pm 13.04$mg Phe per serving for 90% of foods).

The above mathematical procedure is contrasted against a machine learning approach that uses the data in an existing database as training data to infer the Phe in any given food. Specifically, we use the K-Nearest Neighbors (K-NN) classification method using a feature vector containing the (rounded) nutrient data. In other words, the Phe content of the test food is a weighted average of the Phe values of the neighbors closest to it using the nutrient values as attributes. A four-fold cross validation is carried out to determine the hyper-parameters and the training is performed using the United States Department of Agriculture (USDA) food nutrient database. Our tests indicate that this approach is not very accurate for general foods (error $\leq \pm 50$mg Phe per 100g in about 38% of the foods tested). However, for low-protein foods which are typically consumed by PKU patients, the accuracy increases significantly (error $\leq \pm 50$mg Phe per 100g in over 77% foods).

The machine learning approach is more user-friendly than the mathematical approach. It is convenient, fast and easy to use as it takes into account just the nutrient information. In contrast, the mathematical method additionally takes as input a detailed ingredient list, which is cumbersome to be located in a food database and entered as input. However, the Mathematical method has the added advantage of providing error bounds for the Phe estimate. It is also more accurate than the ML method. This may be due to the fact that for the ML method, the nutrition facts alone are not sufficient to estimate Phe and that additional information like the ingredients list is required.

1. INTRODUCTION

Many patients diagnosed with metabolic disorders like Phenylketonuria (PKU) are instructed to follow a strict diet limited in certain nutrients as part of their treatment. In order to do so, patients must know the quantities of these nutrients in all the foods they consume. While the Nutrition Facts Label of commercial foods lists the various nutrients present, it is not a very comprehensive list. Information like amino acids is missing. Also, the nutrients labeled are rounded to the nearest integer. This lack of precision is often a challenge to patients monitoring very strict diets.

The aim of our research is to formulate, implement and compare techniques to estimate missing nutrient quantities present in a food. We believe that the implementation of such techniques in web or phone applications would be useful for patients suffering from metabolic disorders of all kinds to be able to monitor their dietary intake.

Our particular interest in this thesis is with respect to estimation of the content of the amino acid Phenylalanine (Phe) in commercial foods. The Phe intake must be controlled in the diet of patients diagnosed with Phenylketonuria (PKU). PKU is a metabolic disorder that is caused by mutations in the Phenylalanine Hydroxylase (PAH) gene. This in turn affects the secretion of the Phenylalanine Hydroxylase enzyme, which is very important to break down the amino acid, Phe. Mutations in this gene, thus, result in a build up of the Phe content in the body of the patients. An excessive quantity of Phe causes symptoms like intellectual disability, mental disorders, seizures and behavioral problems among others. By taking enormous care to limit their intake of Phe, PKU patients can avoid these adverse effects.

Our aim is to propose techniques to estimate latent quantities from the information present in the Nutrition Facts Label and ordered Ingredients List. We have discussed two approaches for the same.

The first is a mathematical process comprising three steps, as discussed in Chapter 2. The first step considers the protein content in foods and the Phe:protein multipliers to determine Phe content in foods. The revised Phe:protein multipliers discussed in [1] are used for the same to ensure better prediction accuracy of Phe estimates. The protein content taken from the Nutrition Facts Label is a rounded value. This rounded value is used to determine the minimum (min) and maximum (max) protein content in the foods. Subsequently, the min and max protein contents are used along with multipliers [1] to estimate a range for Phe. This step alone is sufficient to identify foods with very high Phe content (like aspartame containing foods). The second step of the Mathematical process is more elaborate in that it considers the Ingredients list information in addition to the Nutrition Facts Label. We list all the Phe:protein multipliers of all the ingredients [2]. The min and max of these multipliers are respectively multiplied with the min and max value of protein (as used in Step 1). This yields a second range of Phe estimates. The third step is a numerical optimization based method. It sets up inequality constraints using the information from the Nutrition Facts label and an ordered list of ingredients written in decreasing order of their weights. This approach is developed on the assumption that no part of any ingredient is dismissed in the preparation process. These inequalities are optimized in two phases. The first phase is an inverse recipe method, where the nutrient content is estimated by approximating ingredients amounts. The second phase comprises refining the above estimates using the simplex algorithm. So, this gives a third range for the Phe estimate. The third step has been discussed in detail in [3–5]. By combining the intervals for Phe obtained in the three steps described above, we obtain a refined range for Phe. By doing so, one achieves better accuracy for Phe estimation as compared to using the computationally intensive third step alone. These results have been published in [6].

Chapter 3 discusses the second approach, which is a Machine Learning (ML) based method. The aim is to use only the nutrition facts and attempt to estimate Phe through a training approach. It makes use of eight nutrients (per 100g) for foods taken from [7] and a K-NN based classification to estimate Phe. The nutrient facts are collected in a feature vector that is used to represent a food. A four-fold cross validation is performed using the data [7] and the algorithm is trained to evaluate a good choice for K (number of nearest neighbors). Once set, this value of K is used for testing. The Phe of a test sample is estimated as the weighted average of Phe values of the K nearest neighbors. The errors in prediction are studied using histograms to analyse the accuracy of the approach.

Finally, in Chapter 4 we compare the two approaches discussed above. This is done by using both the approaches to estimate Phe in a set of 20 commercial foods used in [3, 4]. The errors in estimation of Phe by both the approaches are compared with ground truth from standard database references. Subsequently, the number of foods with Phe estimates from the ML approach lying in the range of Phe predicted by the Mathematical approach gives an indication of accuracy and concurrence of ML approach to the Mathematical method.

There is an increasing focus on employing mobile technology or wearable handheld devices to monitor dietary nutrient intake [8–10]. For example, studies have shown the usefulness of using mobile applications to self monitor one's diet for weight management [11, 12]. Food personalization frameworks are developed with intent of providing a personalized diet [13]. With the growing awareness about health and need for a balanced diet, mobile apps that use real-time questionnaires to give indications about particular foods are common. Such apps provide individualized nutritional recommendations, both to healthy individuals looking for a balanced diet and for patients suffering from pathological conditions. These help combat chronic diet related ailments [14, 15]. To make the system more user-friendly, there are apps that use image segmentation technologies to gather information about the food intake to monitor one's diet [16]. Using advanced computer vision techniques, apps to estimate specific nutrients (like Carbohydrate) from food images can be developed [17]. There has been exemplary work done towards food recognition algorithms for dietary intake management [18, 19]. Our objective is to develop techniques that would be amenable for implementation as a phone or web application. This is important as it would serve as a useful calculator for the PKU patients to estimate their dietary Phe intake. The three methods of the Mathematical approach have been implemented as web applications and are freely available at `https://engineering.purdue.edu/brl/PKU/`.

Although the focus of this study has been towards PKU patients and estimation of Phe, we strongly believe that these techniques can suitably be adopted for estimation of other nutrients (like Lysine) for the treatment of other metabolic disorders.

2. A 3-STEP MATHEMATICAL METHOD TO ESTIMATE PHE

This work has been published in IEEE Access, "A 3-step process to estimate phenylalanine in commercial foods for PKU management" [6].

2.1 Introduction

The mathematical approach described in this chapter is a 3-step method. The first step is based on multipliers suggested by Kim and Boutin in [1]. The second step, although not formally published, was suggested by the same authors. The third step was described and published in [4]. Our contribution is the combination of these three steps. More specifically, we intersect the results of the three steps to obtain a refined range for Phe estimates. As we show in this chapter, our proposed three-step method, which combines the results of the individual methods, yields better estimates than any of the methods applied alone.

In the related web and phone applications, the results for this approach have previously been presented to users as an interval. Specifically, the output of the computation was given as the minimum and the maximum values of the Phe in one serving. However, from a user's viewpoint, we see that values expressed as an estimate of the Phe content $\pm$ some error gives a better idea about the food. Such an estimate (determined from the mid point values of the (min,max) range) $\pm$ some error (max value of Phe - mid point estimate) is the new form of expression of the results which we propose.

As each consecutive method takes as input more information than the previous one, the ranges of possible Phe values tend to become narrower with each consecutive method. We combine the results of the three methods by taking the intersection of all their ranges to produce the final estimate. Note that, as a user progresses through the methods, they are required to enter more and more information in the application. The third method can be perceived as particularly tedious. However, depending on the precision required, the user may choose to stop after the first or second method to gain a fair estimate of the Phe content. For example, such early terminations can be useful to eliminate foods with high Phe content, like foods containing aspartame.

2.2 Three-Step Methodology

2.2.1 Step 1: Phe from Protein Estimation

Step 1 takes as input the rounded protein content from the food label and determines whether the food contains aspartame. As nearly half the weight of aspartame is Phe, this sweetener is generally avoided in the PKU diet. The Phe estimate at this step is obtained by multiplying the minimum and maximum protein content by the appropriate minimum and maximum Phe:protein ratio, respectively, in order to obtain a minimum and maximum Phe amount.

The first step is useful when the user has limited information regarding the ingredients in the food as it considers only the rounded protein content from the food label. If the user is completely unaware of the ingredients, or is not sure if the food contains aspartame, then Phe:protein ratio for aspartame, namely 547 mg Phe per gram protein [2] is used. This gives a very high value for Phe and thus rejects the food as unsuitable for the diet. If the user is certain that the food contains no aspartame, we use the minimum and maximum Phe:protein ratios suggested in [1], namely 20mg and 64.5mg of Phe per gram protein. An optimal refinement can be obtained if more information about the ingredients is known. Specifically, if the food has only fruit based ingredients, the minimum and maximum Phe:protein ratios suggested in [1], namely 20mg and 39mg of Phe per gram protein are used.

To be more precise, let us now explain the method in mathematical terms. Let p be the rounded protein value and let Δ be the maximum rounding error. For example, if the label of a food sold in the US states that it contains 1g of protein, then $p = 1$ and $\Delta = 0.5$. Let $minprotein = p - \Delta$ and $maxprotein = p + \Delta$. Let $minphetoprotein = 20$ and $maxphetoprotein = 64.5$. If the food is known to be made only of fruit ingredients and Phe-free ingredients, then replace the value of $maxphetoprotein$ by 39. If the food contains aspartame (or if it is not know whether it does), replace the value of $maxphetoprotein$ by 547. The minimum and maximum Phe values for the first step are then set to

$$minphe_1 = (minphetoprotein) \times (minprotein),$$

$$maxphe_1 = (maxphetoprotein) \times (maxprotein).$$

If $minphe_1$ is high considering the individual's personal Phe tolerance, the user is advised not to consume the food and the process is terminated. For example, for classical PKU patients whose daily Phe allowance is below $400mg$, a minimum Phe value of 100mg should be ground for dismissing the food. Likewise, if the food contains aspartame (or if it is not known whether it does), the user is advised not to consume the food and the process is terminated.

The Phe estimate for Step 1 is taken to be the middle point of the interval $[minphe_1, maxphe_1]$, and the error of that estimate is set to $\frac{maxphe_1 - minphe_1}{2}$. If the size of the error is considered to be small enough, the user may choose to terminate the process and use the estimate of Step 1 in their diet records. For example, considering the precision of the Phe values obtained by laboratory measurements and the many possible causes of individual food variations, an error value below about $10 - 15mg$ may be considered acceptable.

Observe that the more precise the protein value, the smaller the error of the Phe estimate. When the protein content is rounded to the nearest 0.1g (e.g., for some imported foods sold in the US), the estimate provided is quite accurate. However, Nutrition Facts Labels in the US give the protein content rounded to the nearest 1g. For general foods without aspartame, the smallest maximal error one can obtain is for foods with 0g of protein ($\pm 16.13mg$ Phe). For foods made of fruit-based ingredients, that maximal error decreases to a mere $\pm 9.8mg$ Phe. However, the size of the maximal error grows with the protein content. Thus for US foods whose protein content is 1g or more, this initial step only provides a rough range of possible Phe values and thus is mostly used to quickly screen for foods that are obviously too high in Phe for the patient based on their individual tolerance.

2.2.2 Step 2: Phe from Protein and Ingredient Estimation

The second step takes as input the previously mentioned protein content p and maximum rounding error Δ as well as the ingredient list. Let n be the number of ingredients in the list, and let $phetoprot_i$ be the Phe:protein ratio for ingredient i (from this Phe:protein database [2] or some other database). For this step, the ingredients do not need to be in any particular order. We consider the maximum and minimum Phe:protein ratio for all the ingredients:

$$minphetoprotein = \min\{phetoprot_i\}_{i=1}^{n}$$

$$maxphetoprotein = \max\{phetoprot_i\}_{i=1}^{n}$$

If more than one possibility for an ingredient is found in the Phe:protein database, and thus the phe:protein value is unclear, all values are added to the set before picking the maximum and the minimum. If an ingredient does not contain protein (or only traces of it), or if a minuscule amount of the ingredient is used in the food, then it may be discarded from the list.

Again, we let $minprotein = p - \Delta$ and $maxprotein = p + \Delta$, and the minimum and maximum Phe values for the second step are set to

$$minphe_2 = (minphetoprotein) \times (minprotein),$$

$$maxphe_2 = (maxphetoprotein) \times (maxprotein).$$

The Phe estimate for Step 2 is taken to be the middle point of the interval $[minphe_2, maxphe_2]$, and the error of that estimate is set to $\frac{maxphe_2 - minphe_2}{2}$. If the size of the error is considered to be small enough, the user may choose to terminate the process and use the estimate of Step 2 in their diet records. Note that the estimate of Step 2 should be more accurate (smaller error) than the estimate of Step 1.

2.2.3 Step 3: Numerical Optimization and Interval Intersection

The third step uses the ingredient list and the Nutrition Fact Label. This information is used to set up a set of inequalities which are then solved in order to find the values of $minphe_3$ and $maxphe_3$ using a third method for Phe estimation. The corresponding Phe interval is then intersected with that of Step 2 and 1 in order to produce the final estimate.

To apply the third method of Phe estimation, the ingredients must be listed in decreasing order of weight. This gives us a set of inequality constraints. The method also assumes that there is no loss during the preparation process (e.g., nothing is discarded). This gives us two equality constraints: the sum of each ingredient content equals to a serving size and the weighted sum of a nutrient content for one gram of each ingredient equals to the nutrient content for a serving size. We further consider inequality constraints obtained from the Nutrition Facts Label. The proposed method is applicable even if the nutrient content of some of the ingredients is not fully known. But, in general, the more nutrient information is known, the better the accuracy of the final estimate. Step 3 is performed using six nutrients (protein, sodium, calories, carbohydrates, fat and cholesterol) This Phe estimation method proceeds in two phases which are described in [4].

2.3 Results

We estimated the Phe of 20 commercial foods using our proposed three step Mathematical method. None of the foods chosen contains aspartame, and none of them is made solely of fruit-based ingredients. Details of our data are available at [20].

We used the protein content rounded to the nearest gram (± 0.5g error) in order to show the accuracy one would expect when using US food labels and the protein content rounded to 0.1g (± 0.05g error) to incorporate food labels of products from non-US countries.

Tables 2.1 and 2.3 show the Phe estimation results obtained from Step 1 for the 1g and 0.1g protein rounding respectively. Tables 2.2 and 2.4 show the Phe estimation results obtained from Step 2 and Step 3, along with the final estimates for the 1g and 0.1g protein rounding respectively. For each step, the results comprises the Min Phe (in mg), Max Phe (in mg), estimated Phe (in mg) and the error (in mg). Tables 2.1 and 2.3 also contain the protein content of the foods. Tables 2.2 and 2.4 show the final Phe estimate with error that is obtained by intersecting the (min,max) Phe intervals of all the three steps.

From Table 2.1 we note that the error obtained for foods with 0g protein is only about 16mg as seen in 8 out of the 20 foods when using protein content rounded to nearest gram. If the protein content is rounded to nearest 0.1g, then the error is 10 times smaller as seen in Table 2.3. With increasing protein content, the error increases. This can be observed in Table 2.1 wherein error increases from 43mg to 88mg when protein content increases from 1g per serving to 3g per serving. A similar trend is seen in Table 2.3 (with rounding of 0.1g) wherein error increases from 24mg to 64mg when protein content per serving increases from 1g to 3g. Foods containing a very high Phe content can be identified using Step 1 alone and the algorithm can be terminated. For example, "Yoplait Original Strawberry" contains a min Phe of 110mg (with 0.5g rounding error) and 115mg (with 0.05g rounding error). This cell has been marked yellow in Tables 2.1 and 2.3 respectively. Such high-Phe content foods can be rejected as unsuitable for a classic PKU diet.

Performing Step 2 improves the Phe estimates. As seen from Table 2.2, the errors reduce in all the cases compared to errors obtained in Step 1. The accuracy for Step 2 depends on the spread of the Phe:protein ratios for the ingredients. The smaller the range of (Min,Max) of the ratios, the larger the improvement in accuracy seen from

Step 1. As seen for Food # 1 in Tables 2.1 and 2.2, the error reduces from 43mg to 30mg.

Step 3 further improves the accuracy by lowering the errors obtained. The error for Food # 1 reduces to 16mg after performing Step 3. Although Step 3 takes a lot more input data, it is computationally expensive and yields the lowest errors among the three steps. A similar trend of improvement in accuracies can be expected with Steps 2 and 3 for the 0.05g protein rounding error case.

Subsequently, by combining the results of the three steps and intersecting the intervals, we get better results. For example, the error for Food # 1 reduces to 13mg by combining intervals. This is possible since the intervals obtained from the 3 steps are not necessarily nested. Thus, it leads to an overall more refined estimate with lower errors than those achieved by using the methods individually. Such an improvement is seen in 3 foods (Foods # 1,3,7) for the 0.5g precision case and in 7 foods (Foods # 1,2,3,7,11,14,20) for the 0.05g precision case as seen from the yellow shaded cells in Tables 2.2 and 2.4 respectively. An important observation is that by increasing the precision of the input values, we can achieve much smaller errors than before. As seen in Table 2.4, the errors after final intersection are very small compared to final errors in Table 2.2.

Table 2.1.
Phenylalanine Content Estimate After Step 1 with Protein Content Precision of ± 0.5g.

Food Number #	**Description (serving size)**	**Protein Content (in g)**	**Min Phe (in mg)**	**Max Phe (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**
1	**Carr's Whole Wheat Crackers (17 g)**	1	10	96.75	53.38	43.38
2	**Heinz Tomato Ketchup (17 g)**	0	0	32.25	16.13	16.13
3	**KIT KAT Milk Chocolate (42 g)**	3	50	225.75	137.88	87.88
4	**Campbell's Tomato soup (122 g)**	2	30	161.25	95.63	65.63
5	**Cheerios Cereal (28 g)**	3	50	225.75	137.88	87.88
6	**Rice Krispies Cereal (33 g)**	2	30	161.25	95.63	65.63
7	**Enchilada Sauce (60 g)**	1	10	96.75	53.38	43.38
8	**Eggo waffle (70 g)**	4	70	290.25	180.13	110.13
9	**Garlic chili pepper sauce (9 g)**	0	0	32.25	16.13	16.13
10	**Salsa sauce (30 g)**	0	0	32.25	16.13	16.13
11	**Simply potatoes Garlic mashed potatoes (124 g)**	3	50	225.75	137.88	87.88
12	**Butter with Canola Oil (14 g)**	0	0	32.25	16.13	16.13
13	**Go-Gurt (64 g)**	2	30	161.25	95.63	65.63
14	**Jell-O Gelatin Snacks-Strawberry (98 g)**	1	10	96.75	53.38	43.38
15	**Ore-Ida French fries (84 g)**	2	30	161.25	95.63	65.63
16	**Spicy Brown Mustard (5 g)**	0	0	32.25	16.13	16.13
17	**Starburst Fruit Chews (40 g)**	0	0	32.25	16.13	16.13
18	**Vinaigrette Balsamic Dressing (31 g)**	0	0	32.25	16.13	16.13
19	**Yoplait Original Strawberry (170 g)**	6	110	419.25	264.63	154.63
20	**ALTOIDS peppermint (2 g)**	0	0	32.25	16.13	16.13

Table 2.2.
Phenylalanine Content Estimate After Steps 2 and 3 with Protein Content Precision of ±0.5g.

#	Step 2				Step 3[1]				Final Intersection	
	Min Phe (in mg)	**Max Phe (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**	**Min Phe (in mg)**	**Max Phe (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**
1	20.55	79.69	50.12	29.57	53.61	85.11	69.36	15.75	66.65	13.04
2	0.00	32.14	16.07	16.07	1.20	6.57	3.89	2.69	3.89	2.69
3	87.72	185.94	136.83	49.11	144.27	191.53	167.90	23.63	165.11	20.84
4	30.91	132.81	81.86	50.95	40.69	95.45	68.07	27.38	68.07	27.38
5	120.72	199.23	159.97	39.26	179.86	180.51	180.19	0.32	180.19	0.32
6	78.87	134.62	106.74	27.87	91.54	94.80	93.17	1.63	93.17	1.63
7	12.20	96.43	54.31	42.12	0.41	34.14	17.28	16.87	23.17	10.97
8	143.82	297.25	220.53	76.71	196.26	216.35	206.31	10.05	206.31	10.05
9	0.00	16.58	8.29	8.29	2.65	5.27	3.96	1.31	3.96	1.31
10	0.00	26.73	13.37	13.37	7.90	18.23	13.07	5.17	13.07	5.17
11	70.31	183.33	126.82	56.51	139.51	162.23	150.87	11.36	150.87	11.36
12	0.00	26.19	13.10	13.10	12.06	17.66	14.86	2.80	14.86	2.80
13	31.07	129.55	80.31	49.24	116.38	120.95	118.67	2.29	118.67	2.29
14	10.00	51.00	30.50	20.50	10.01	30.44	20.23	10.22	20.23	10.22
15	40.35	160.72	100.53	60.19	77.64	78.76	78.20	0.56	78.20	0.56
16	0.00	32.14	16.07	16.07	10.11	10.16	10.14	0.03	10.14	0.03
17	0.00	18.00	9.00	9.00	0.00	4.48	2.24	2.24	2.24	2.24
18	0.00	32.14	16.07	16.07	0.00	5.53	2.77	2.77	2.77	2.77
19	113.92	336.82	225.37	111.45	287.11	291.08	289.10	1.98	289.10	1.98
20	0.00	10.36	5.18	5.18	0.43	4.22	2.33	1.90	2.33	1.90

[1]The results for Step 3 have been taken from [4].

Table 2.3.
Phenylalanine Content Estimate After Step 1 with Protein Content Precision of ± 0.05g.

Food Number #	**Description (serving size)**	**Protein Content (in g)**	**Min Phe (in mg)**	**Max Phe (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**
1	**Carr's Whole Wheat Crackers (17 g)**	1.0[1]	19.00	67.73	43.36	24.36
2	**Heinz Tomato Ketchup (17 g)**	0.2	3.00	16.13	9.56	6.56
3	**KIT KAT Milk Chocolate (42 g)**	2.8	55.00	183.83	119.41	64.41
4	**Campbell's Tomato soup (122 g)**	1.8	35.00	119.33	77.16	42.16
5	**Cheerios Cereal (28 g)**	3.4	67.00	222.53	144.76	77.76
6	**Rice Krispies Cereal (33 g)**	2.2	43.00	145.13	94.06	51.06
7	**Enchilada Sauce (60 g)**	1.0[1]	19.00	67.73	43.36	24.36
8	**Eggo waffle (70 g)**	3.9	77.00	254.78	165.89	88.89
9	**Garlic chili pepper sauce (9 g)**	0.0[1]	0.00	3.23	1.61	1.61
10	**Salsa sauce (30 g)**	0.0[1]	0.00	3.23	1.61	1.61
11	**Simply potatoes Garlic mashed potatoes (124 g)**	2.8	55.00	183.83	119.41	64.41
12	**Butter with Canola Oil (14 g)**	0.0[1]	0.00	3.23	1.61	1.61
13	**Go-Gurt (64 g)**	2.4	47.00	158.03	102.51	55.51
14	**Jell-O Gelatin Snacks-Strawberry (98 g)**	1.0	19.00	67.73	43.36	24.36
15	**Ore-Ida French fries (84 g)**	2.0[1]	39.00	132.23	85.61	46.61
16	**Spicy Brown Mustard (5 g)**	0.2	3.00	16.13	9.56	6.56
17	**Starburst Fruit Chews (40 g)**	0.0	0.00	3.23	1.61	1.61
18	**Vinaigrette Balsamic Dressing (31 g)**	0.0[1]	0.00	3.23	1.61	1.61
19	**Yoplait Original Strawberry (170 g)**	5.8	115.00	377.33	246.16	131.16
20	**ALTOIDS peppermint (2 g)**	0.0	0.00	3.23	1.61	1.61

[1] Exact values not found. Rounded values with increased precision of 0.1g considered.

Table 2.4.
Phenylalanine Content Estimate After Steps 2 and 3 with Protein Content Precision of ±0.05g.

#	Step 2				Step 3 [1] [4]				Final Intersection	
	Min Phe (in mg)	**Max Phe (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**	**Min Phe (in mg)**	**Max Phe (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**	**Phe estimate (in mg)**	**Error (in mg)**
1[2]	39.04	55.78	47.41	8.37	53.61	85.11	69.36	15.75	54.70	1.08
2	3.09	16.07	9.58	6.49	1.20	6.57	3.89	2.69	4.83	1.74
3	96.49	151.41	123.95	27.46	144.27	191.53	167.90	23.63	147.84	3.57
4	36.06	98.28	67.17	31.11	40.69	95.45	68.07	27.38	68.07	27.38
5	161.76	196.38	179.07	17.31	179.86	180.51	180.19	0.32	180.19	0.32
6	113.04	121.15	117.10	4.06	91.54	94.80	93.17	1.63	-[3]	-[3]
7[2]	23.17	67.50	45.34	22.17	0.41	34.14	17.28	16.87	28.66	5.48
8	158.20	260.92	209.356	51.36	196.26	216.35	206.31	10.05	206.31	10.04
9[2]	0.00	1.66	0.83	0.83	2.65	5.27	3.96	1.31	-[3]	-[3]
10[2]	0.00	2.67	1.34	1.34	7.90	18.23	13.07	5.17	-[3]	-[3]
11	77.34	149.29	113.31	35.97	139.51	162.23	150.87	11.36	144.40	4.89
12[2]	0.00	2.62	1.31	1.31	12.06	17.66	14.86	2.80	-[3]	-[3]
13	48.68	126.95	87.82	39.14	116.38	120.95	118.67	2.29	118.67	2.28
14	19.00	35.70	27.35	8.35	10.01	30.44	20.23	10.22	24.72	5.72
15[2]	52.45	131.79	92.12	39.67	77.64	78.76	78.20	0.56	78.20	0.56
16	4.03	16.07	10.05	6.02	10.11	10.16	10.14	0.03	10.14	0.02
17	0.00	1.80	0.90	0.90	0.00	4.48	2.24	2.24	0.90	0.90
18[2]	0.00	3.21	1.61	1.61	0.00	5.53	2.77	2.77	1.61	1.60
19	119.10	303.14	211.12	92.02	287.11	291.08	289.10	1.98	289.10	1.98
20	0.00	1.04	0.52	0.52	0.43	4.22	2.33	1.90	0.74	0.3

[1] The results for Step 3 have been taken from [4].

[2] Exact values not found. Rounded values with increased precision of 0.1g considered.

[3] Feasible final intervals cannot be determined.

[4] Computed using $\pm0.5g$ protein precision instead of $\pm0.05g$.

2.4 Web Implementation

As discussed in Chapter 1, the aim of the project has been to develop techniques for Phe estimation that are user friendly and amenable to implementation as smart phone or web applications. The goal is to aid PKU patients make appropriate food choices and monitor their dietary Phe intake.

A prototype of such a web application has been implemented for the three separate methods discussed in this chapter. It is readily available at `https://engineering.purdue.edu/brl/PKU/`.

The "protein multiplier method" is an implementation of the first step. As shown in Figure 2.2, it takes as input only the protein content. It also ensures whether the food contains the high Phe ingredient aspartame. It enquires if the food contains only fruits and Phe-free ingredients so as to choose the right Phe:protein multipliers. As shown in Figure 2.3, the result is expressed both as range of (min,max) Phe values and as a Phe estimate with an ±error.

The "Phe:protein ratio ingredients method" implements the second step. In addition to the protein content, it takes as input the ingredients list as shown in Figure 2.4. The web app takes into account the values for Phe:Protein ratios of ingredients of various food products from standard databases [2]. These databases are used as a look-up table to search for the values required in Step 2. The result is expressed as range of (min,max) Phe and as (Phe estimate±error) as seen in Figure 2.5.

Finally the "Inverse recipe method" is an implementation of the third step. It takes as input, the serving size, ingredients list and nutrient information (Figure 2.6). The results of the first phase include the approximate ingredients amounts and the corresponding Phe estimates (result expressed in both forms) as shown in Figure 2.7. The second phase results show the maximizers and minimizers from the Simplex method and the final refined Phe estimates (result expressed in both forms) as shown in Figure 2.8.

The web page is developed using HTML, PHP and Python. The first two steps and the first phase of the third step were implemented by Jieun Kim. We implemented the second phase (Simplex) of the third step. We also expressed the results in the form of (phe estimate±error) for all three steps, which provides more readability than the previous (Min,Max) Phe representation.

Currently, the app requires the user to manually enter the ingredients, answer questions about the presence of aspartame and indicate if the food is solely fruit-based. However, its implementation can be further enhanced to include OCR (Optical Character Recognition) techniques to read the values of protein content and the ingredient composition directly from the scanned images of the food label taken by a smart-phone to make it more user-friendly. Another option would be to read the bar code. One could also extend the scope of the app for estimation of other amino acids so as to extend its application to other inborn metabolic disorders.

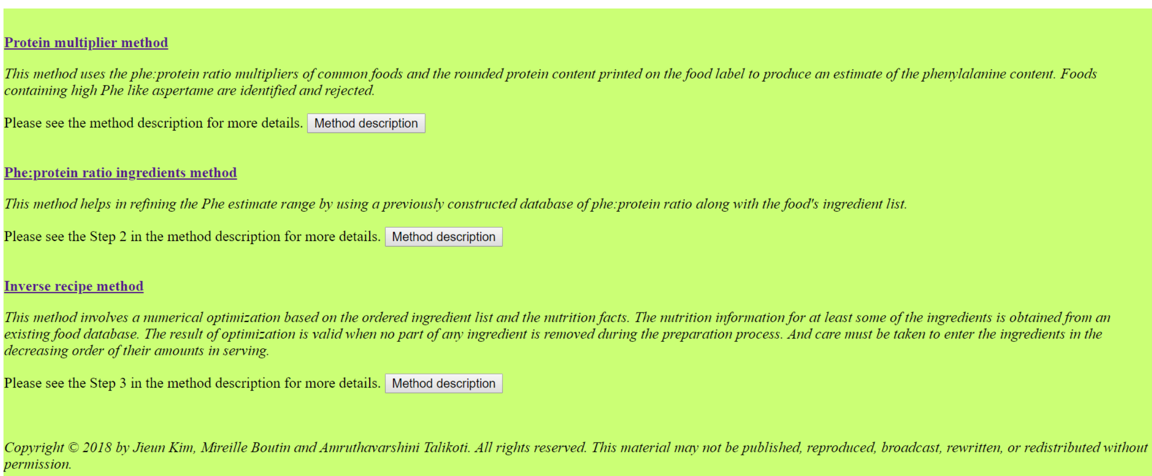


Fig. 2.1. Home Page of the Web Application

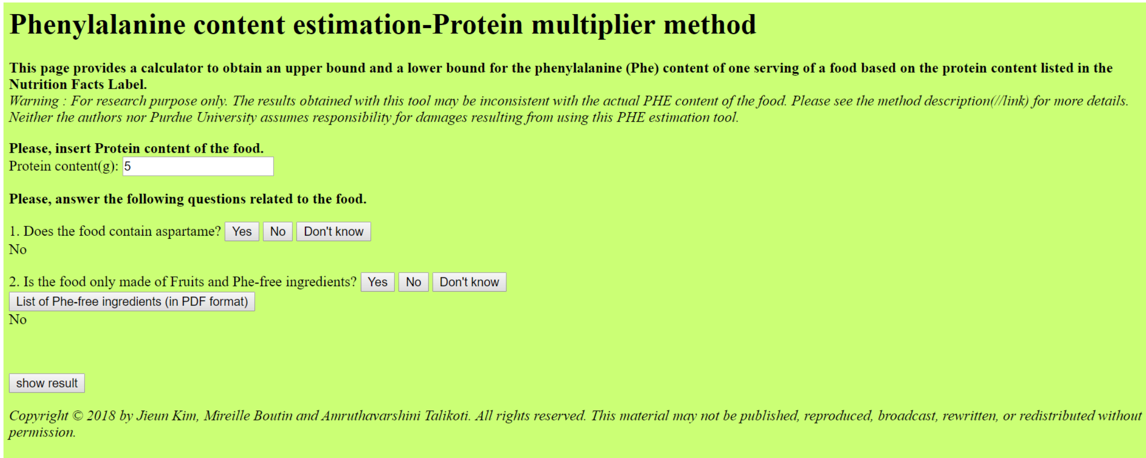


Fig. 2.2. Input Data for Method 1

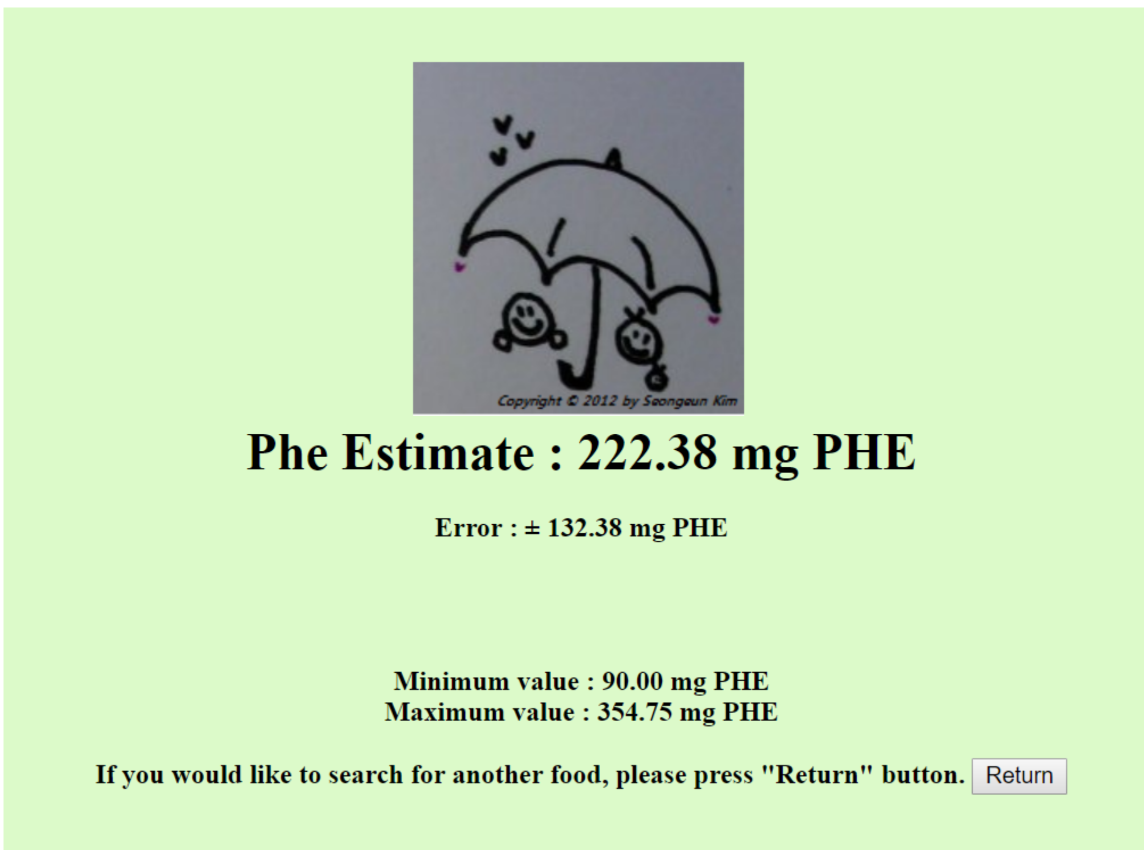


Fig. 2.3. Results of Method 1

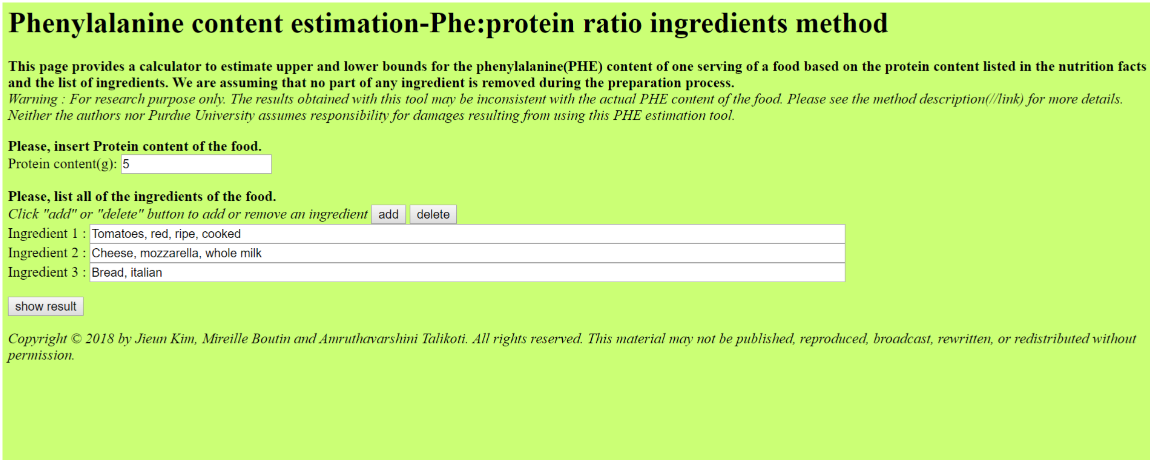


Fig. 2.4. Input Data for Method 2

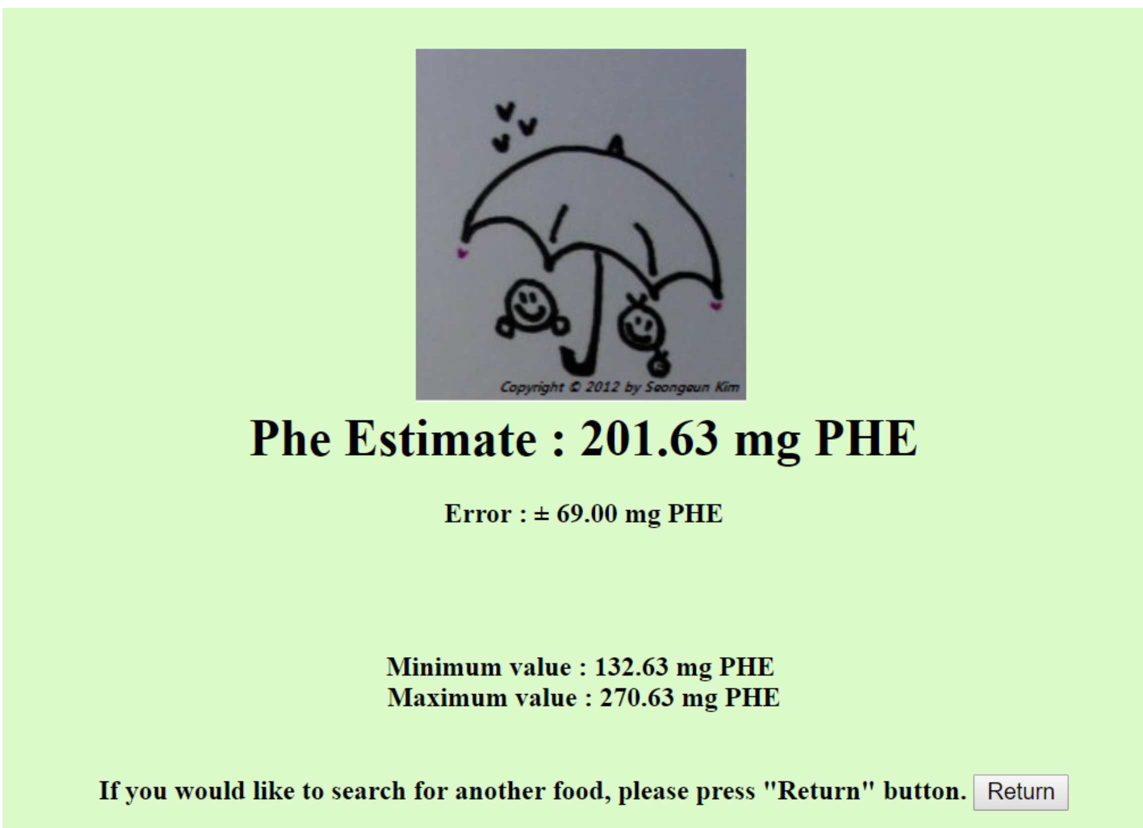


Fig. 2.5. Results of Method 2

Fig. 2.6. Input Data for Method 3

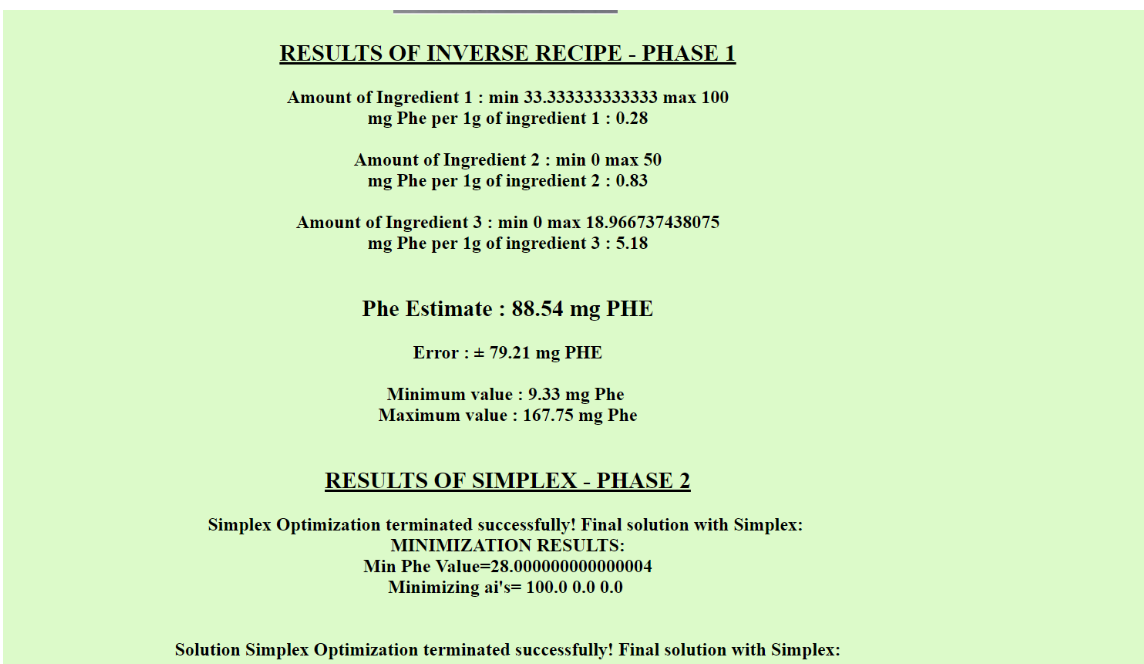


Fig. 2.7. Results of Method 3

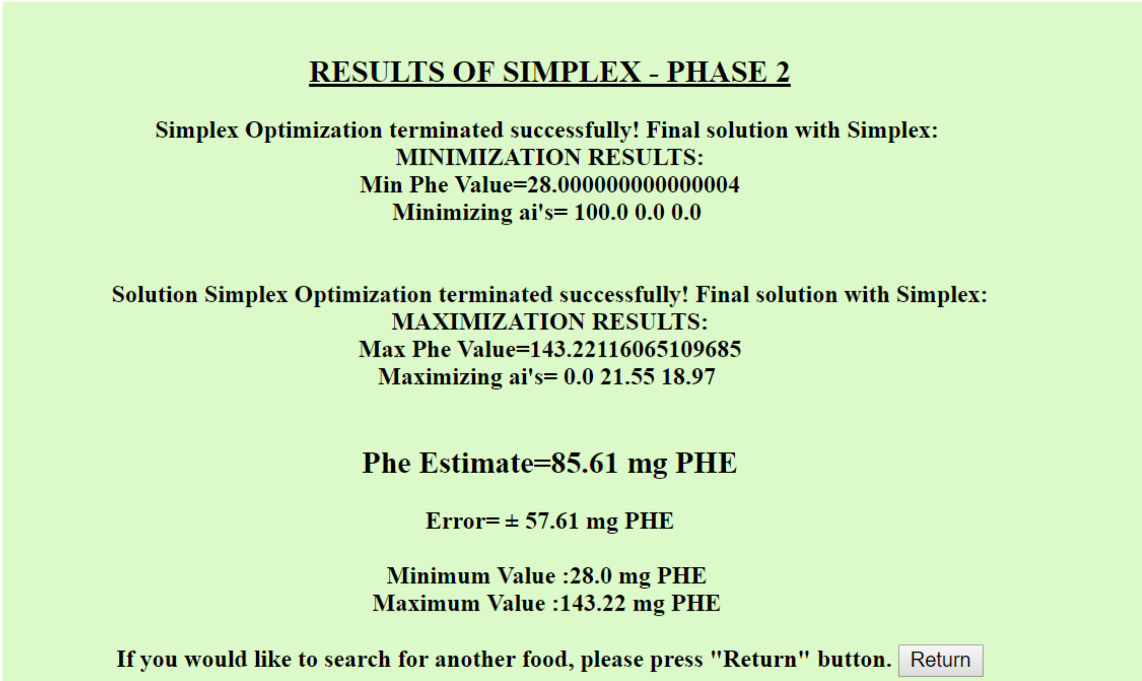


Fig. 2.8. Results of Method 3 Continued

2.5 Conclusion

Some people may require information about commercial foods they consume that may not be readily available on the Nutrition Facts Label and ordered ingredients list. This information (with good precision) may be crucial for dietary management of metabolic disorders like PKU. To combat this issue, a 3-step mathematical method was proposed that determines latent values of nutrients (like Phe) from the data available on the label.

The third step is based on the assumption that no ingredients are missed in the preparation process. The first two methods hold no such requirements. The overall method is applicable even if the nutrient content of some ingredients is unknown. In our experiments, our method was shown to work well, with an error less than $\pm 13mg$ for 18 out of the 20 foods assuming protein values rounded to nearest gram.

By adding one digit to the protein content precision, the accuracy further improves, as seen in Table 2.4. However, in some cases, the Phe interval after the first two steps is inconsistent with the one obtained after the third step. This is likely due to ingredient loss during the preparation process. Note that this happened in 4 out of the 20 foods considered. However, the error in Phe in these 4 cases is very small (less than 4-5mg), and so there is no need to improve it. In the remaining 16 out of the 20 foods, intersecting the intervals yield a non-empty Phe interval. For those foods, the error is less than $\pm 6mg$ for 14 out of the 16 foods.

Each subsequent step of the 3-step process takes more input data, but yields more refined Phe estimates. This in turn leads to lower errors with each step performed. Also, since the intervals produced by the three steps are not nested necessarily, better results can be obtained by combining the results. Considering the good accuracy of our results, and the facts that the method provides clear error bounds on the Phe estimate, we believe that this mathematical method can serve as a useful tool for PKU management. It would be interesting to extend this work to other nutrients so as to extend its application to other metabolic disorders.

3. A MACHINE LEARNING BASED METHOD TO ESTIMATE PHE

This approach is based on the intuition that nutrient facts are related to the amount of Phe in the food [1]. The idea to use K-Nearest Neighbors approach for Phe estimation was originally suggested by J. Kim. The realization of the methodology, implementation and results constitute our contribution to the work.

3.1 Introduction to Data and Pre-processing

The objective of the Machine Learning approach is to estimate Phe from the Nutrition Facts Label. To develop this approach, we used data from the USDA food nutrient database [7] to gain relevant nutrient information of various foods. This database provides 5079 foods with suggested values for Phe content. While it is important to look at all foods, for our application, we are particularly interested in low protein and low Phe foods.

We consider eight nutrition facts which include Protein (in g), Total Lipids or Fats (in g), Carbohydrates by difference (in g), Energy (in kcal), Total Sugars (in g), Total Dietary Fibers (in g), Sodium (in mg) and Cholesterol (in mg). These nutrition facts, which are values per 100g of the food, are placed in a feature vector. Subsequently, these feature vectors are used to estimate the Phe content (in g) per 100g of the food.

We whiten the data as follows [21]. This statistical transformation is carried out so that dimensions are made statistically uncorrelated. This is done by ensuring that the data has an identity covariance matrix.

Let us assume a data matrix X of dimensions (k X n), wherein k be the number of attributes/features and n be the number of samples. Each row of this matrix is populated by subtracting from the i-th attribute of the samples, the mean of the i-th attribute of all samples. This can be represented as follows:

$$\begin{array}{c|ccc} & X_1 & \cdots & X_n \\ \hline f_1 & X_1 - m(f_1) & \cdots & X_n - m(f_1) \\ \vdots & \vdots & \ddots & \vdots \\ f_k & X_1 - m(f_k) & \cdots & X_n - m(f_k) \end{array},$$

wherein $X_1 \cdots X_n$ are the n samples which are k-dimensional, $f_1 \cdots f_k$ are the k attributes, $m(f_1) \cdots m(f_k)$ are the means of the samples along the $f_1 \cdots f_k$ dimensions. The covariance of each of the dimensions with respect to each other is given by constructing a covariance matrix Σ as follows:

$$\Sigma = cov(X) = E(XX^T) \approx \frac{XX^T}{n}.$$

As per the above definition, Σ is symmetric and positive semi-definite. So, its Singular Value Decomposition (SVD) is

$$\Sigma = EDE^{-1},$$

wherein E is a (K X K) sized matrix with each column as an eigenvector of Σ, D is a diagonal matrix whose diagonal elements D_{ii} are eigenvalues corresponding to the eigenvectors of the i-th column of E. Transforming Σ into a diagonal matrix D can be done as

$$E^{-1}\Sigma E = D. \tag{3.1}$$

The aim is to transform the data matrix X into a new data matrix Y using a transforming matrix W_D

$$Y = W_D X, \tag{3.2}$$

whose dimensions are uncorrelated. In other words, Y has a diagonal covariance matrix. We want a transformation W_D that makes

$$D = cov(Y) = E(YY^T). \tag{3.3}$$

From equations 3.1 to 3.3, we can derive that

$$W_D = E^T.$$

Now we also need to ensure an identity covariance matrix. This is done by scaling the dimensions which are now uncorrelated. In other words, we need a transformation that makes D, an identity matrix:

$$\begin{aligned} D^{-1}D &= I, \\ D^{-1} &= D^{-1/2} I D^{-1/2}, \\ D^{-1/2}E^{-1}\Sigma E D^{-1/2} &= I. \end{aligned}$$

Let W_W be the whitening matrix that ensures $cov(Y) = I$. This is given by

$$W_W = D^{-1/2}E^T = D^{-1/2}W_D = D^{-1/2}E^T.$$

This whitening matrix is determined using the training data and is used to transform the train, validation and test data before being used.

Also, covariance matrices are computed before and after whitening.

	0	1	2	3	4	5	6	7
0	117.627	0.155581	-94.1483	120.17	-25.0625	-6.36322	91.5667	346.704
1	0.155581	265.506	-37.9319	2187.04	-7.47466	-0.508744	348.796	81.6494
2	-94.1483	-37.9319	526.734	1289.97	94.4631	42.9265	1110.59	-705.43
3	120.17	2187.04	1289.97	24814.9	192.372	110.735	7864.45	-429.106
4	-25.0625	-7.47466	94.4631	192.372	74.8704	4.46937	90.0964	-139.567
5	-6.36322	-0.508744	42.9265	110.735	4.46937	14.1312	17.2927	-87.8201
6	91.5667	348.796	1110.59	7864.45	90.0964	17.2927	158884	-860.429
7	346.704	81.6494	-705.43	-429.106	-139.567	-87.8201	-860.429	16919.4

Fig. 3.1. Covariance Matrix Before Whitening

	0	1	2	3	4	5	6	7
0	1	1.84163e-16	-6.84783e-17	-3.24034e-16	-3.57836e-16	-2.32535e-16	-2.09806e-16	5.90516e-15
1	1.84163e-16	1	2.16217e-16	-7.11592e-16	2.06163e-16	-3.7124e-16	-4.5458e-17	-4.42457e-15
2	-6.84783e-17	2.16217e-16	1	1.20056e-15	-1.47797e-15	-1.02572e-16	-2.56896e-15	-2.57919e-14
3	-3.24034e-16	-7.11592e-16	1.20056e-15	1	2.97225e-17	-1.08283e-15	-1.25417e-15	-2.54991e-14
4	-3.57836e-16	2.06163e-16	-1.47797e-15	2.97225e-17	1	-1.04553e-15	-1.53683e-15	-8.09283e-15
5	-2.32535e-16	-3.7124e-16	-1.02572e-16	-1.08283e-15	-1.04553e-15	1	-1.8766e-15	-2.45648e-14
6	-2.09806e-16	-4.5458e-17	-2.56896e-15	-1.25417e-15	-1.53683e-15	-1.8766e-15	1	-3.77184e-15
7	5.90516e-15	-4.42457e-15	-2.57919e-14	-2.54991e-14	-8.09283e-15	-2.45648e-14	-3.77184e-15	1

Fig. 3.2. Covariance Matrix After Whitening

Fig. 3.1 shows one of the covariance matrices prior to whitening. The corresponding eigenvalues are obtained using SVD as (1.59359821e+05, 2.46326654e+04, 1.69305868e+04, 5.14085932e+02, 1.12549629e+02, 5.52634670e+01, 1.09029561e+01, 1.44379219e+00). Fig. 3.2 shows the corresponding covariance matrix after whitening. The eigenvalues for this obtained using SVD are all unity (1, 1, 1, 1, 1, 1, 1, 1) as expected. This ensures that the data has been correctly whitened. An important observation is that all the eigenvalues even prior to whitening are not even close to zero. This implies that the 8 nutrients considered form 8 dimensional vectors and that any attempt to express these vectors using fewer dimensions using a linear change of basis would result in a significant information loss. Another important observation was regarding the covariance matrix constructed using the 8 nutrients and Phe values, for a total of 9 attributes. (See Figure 3.3). Performing an SVD, the eigenvalues of this matrix are (5.42918107e+05, 2.51725937e+04, 1.77892883e+04, 5.11152935e+02, 1.11335891e+02, 5.32776382e+01, 1.13777216e+01, 1.28348000e+00, 1.12602609e-02). The ratio between the largest and smallest eigenvalue is 10^7. This large ratio suggests a linear dependency between the Phe content and the 8 nutrients in any food considered. This reinforces our motivation that K-NN classification might be a worthwhile attempt at solving the problem of Phe estimation from nutrition facts.

	0	1	2	3	4	5	6	7	8
0	117.174	2.26885	-94.0254	138.813	-24.3817	-6.13618	-67.5366	374.838	4.81667
1	2.26885	263.856	-35.1891	2190.54	-6.43324	-0.20599	264.368	109.803	0.232512
2	-94.0254	-35.1891	524.748	1308.46	88.842	42.5697	914.598	-697.39	-3.14201
3	138.813	2190.54	1308.46	24972.9	182.459	112.97	5724.81	-12.3637	9.49413
4	-24.3817	-6.43324	88.842	182.459	70.8326	4.36259	68.5326	-134.031	-0.931021
5	-6.13618	-0.20599	42.5697	112.97	4.36259	14.5355	-13.2572	-88.5279	-0.126695
6	-67.5366	264.368	914.598	5724.81	68.5326	-13.2572	542853	-314.237	0.0913159
7	374.838	109.803	-697.39	-12.3637	-134.031	-88.5279	-314.237	17751.3	15.3468
8	4.81667	0.232512	-3.14201	9.49413	-0.931021	-0.126695	0.0913159	15.3468	0.211027

Fig. 3.3. Covariance Matrix of 9 Attributes (Including Phe) Before Whitening

3.2 K-NN Method Description

In order to select the appropriate number of neighbors K, a four-fold cross-validation is carried out with K-NN classification with K ranging from 1 to 50. After a random shuffling, the data (5079 foods) is divided into four sets namely A (1270 foods), B (1270 foods), C (1270 foods) and D (1269 foods). In each fold, one of the above sets is treated as test data and the other three sets are combined to be used as the train data. 20% of the train data is set aside as validation data within each fold. For example, for fold 1 of the algorithm, D is used as test data, 20% of (A+B+C) is used as validation data and the remaining 80% of (A+B+C) is the train data.

This step is used to determine a good value for the hyper-parameter, K. Once the value of K is set, it will be used for the rest of the experiments.

Cross-validation is performed to determine the best K by evaluating validation accuracies for different values of K (1 to 50). For each fold, Phe estimates are evaluated for the validation data using the training data in that particular fold for a given value of K. This is done by determining the K nearest neighbors to the validation data sample in the train data. Subsequently, a weighted average of the Phe values of these neighbors is assigned as the Phe estimate to the validation data sample. This is done as follows

$$Phe(\text{validation sample}) = \frac{w_1 Phe(n_1) + w_2 Phe(n_2) + ... + w_K Phe(n_K)}{w_1 + w_2 + ... + w_K},$$
$$w_i = \frac{1}{d_i}, i = 1...K,$$

wherein, $Phe(\text{validation sample})$ is the Phe estimate assigned to the validation sample under consideration, $n_1...n_K$ are the K nearest neighbors to the validation sample, $w_1...w_K$ are the weights assigned to neighbors $n_1...n_K$, $Phe(n_1)...Phe(n_K)$ are the Phe values of the neighbors $n_1...n_K$ and $d_1...d_K$ are the distances from the validation sample under consideration to the neighbors $n_1...n_K$.

Subsequently, the absolute difference between the actual Phe value and the estimated Phe value (using weighted average) for the validation sample is determined. If this difference is less than 0.1g Phe, then the estimation is labelled as accurate for that particular sample. The number of accurately estimated samples divided by the total number of samples in a fold gives the accuracy for that fold for a given K. Such validation accuracies are determined for each fold for all possible values of K ranging from 1 to 50. Fig. 3.4 shows a plot of these accuracies against different values of K for each fold. This is used to determine the best possible values of K to be 4 or 7 that yield the highest validation accuracies.

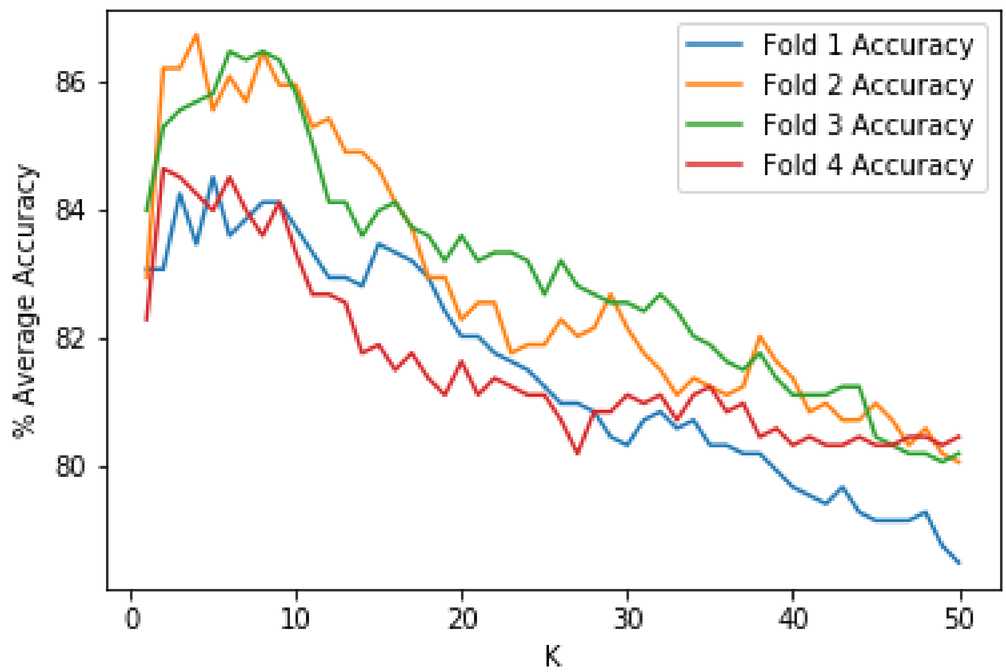


Fig. 3.4. Accuracy Versus K for Each Fold of Validation

Having determined the best values of K from validation, we now move on to testing. The process of testing is carried out similar to validation and Phe estimates of the test samples are calculated using the training data of each fold. Once the Phe estimates for the test samples are determined, a histogram of the errors are plotted for analysis. The error is calculated as the absolute difference between the actual and estimated Phe values. A histogram of these errors are plotted for each of the folds (Figures 3.5 to 3.22). For further analysis, foods estimated with error $\leq$ 50mg Phe per 100g are considered to be accurately predicted and grouped in good accuracy foods. Similarly, foods estimated with errors > 50mg Phe per 100g are grouped under bad accuracy foods.

3.3 Methods to Improve the Choice of Metric

The metric used to find distance to the neighbors in the train set from any sample in the validation or test sets is Euclidean. Various methods to improve the accuracy by altering the distance metric were experimented.

Firstly an exhaustive search was carried out by changing the weights of the Euclidean distance metric as values from (0,0.2,0.4,0.6,0.8,1.0). However, this was not a feasible option because of the huge number of permutations (6^8) and limited time frame.

So, instead, we worked towards a numerical gradient ascent method. The initial solution for weights is considered as

$$w = (1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0).$$

The gradient vector is calculated for each dimension i=1...8 as follows

$$f^{'(w_i)}(w_1, ..., w_i, ...w_8) \approx \frac{f(w_1, ..., w_i + \Delta, ..., w_8) - f(w_1, ..., w_i, ..., w_8)}{\Delta},$$

with $\Delta = 0.01$ The gradient vector is given by

$$\nabla f = (f^{'(w_1)}, ..., f^{'(w_i)}, ...f^{'(w_8)}).$$

The next solution for weights using gradient ascent is given by

$$w' = w + \gamma \nabla f,$$

wherein γ is the learning rate chosen as = 0.05 or 0.1. Using the new set of weights, w' in the distance metric, the new accuracies are determined. However, we see that there is again no significant improvement in the accuracy by performing the gradient ascent. Thus, we can say that we are possibly at a local maxima in the accuracy plot.

3.4 Implementation

The source code is written in Python 2.7 using an open source cross-platform Integrated Development Environment (IDE), Spyder as part of the Anaconda Navigator

Package. The system used for running the code is a Windows 10 Personal Computer (PC) with 64-bit operating system working on Intel(R) Core(TM) i5-7200U CPU. The python code and related files can be found at [22].

3.5 Numerical Experiments

Different metrics were experimented with to represent the accuracy for our proposed KNN Phe estimation method. One such metric represents the percentage error determined as follows

$$\%Error = \frac{\text{Predicted Phe - Actual Phe}}{\text{Actual Phe}}$$

However, the issue with this metric is that it is not a suitable choice to represent error in foods with an actual Phe content of 0g. Hence, the metric finally chosen was a histogram of the absolute value of error (in mg per 100g food) defined as

$$Error = |\text{Predicted Phe - Actual Phe}|$$

These histograms are as shown in Figures 3.5 to 3.22. The x-axis denotes the Error in Phe Estimation (in g) per 100g of food. The y-axis denotes the number of foods with the errors in ranges as shown along the x-axis. All but the last column in the histogram are the foods estimated with what we consider as good accuracy. The last bar contains those foods with poorly predicted Phe estimates (error $>$ 50mg Phe per 100g food). As we can see in the Fig. 3.5 (K=4), the first column, which represents foods predicted with a great accuracy (Error $\leq$ 5mg Phe per 100g food), contains about 70 foods. The last column contains over 700 foods that have not been well estimated for Phe (Error $>$ 50mg Phe per 100g food). We can also observe that the number of foods predicted with a good accuracy (Error $\leq$ 50mg Phe per 100g food) represented by the first 10 columns in the graph is lesser than the number of foods predicted with a bad accuracy and represented in the last column. In fact, on an average, 1972 out of 5079 foods are predicted with a good accuracy and the remaining 3107 foods are predicted with a bad accuracy. In other words, the Phe of

38% of the foods is estimated with a good accuracy, as summarized in Table 3.1. The histograms for the case of K=7 are similar to the results described for K=4 case as can be seen in Fig. 3.6. These results as described in the Figure 3.5 and 3.6 are for the general case which comprises all foods.

As discussed in Section 3.2, the number of accurately predicted foods (foods estimated with error $\leq$ 50 mg Phe) divided by the total number of foods gives the good accuracy of a fold. Similarly, the bad accuracy is computed for each fold (using foods estimated with error $>$ 50 mg Phe). The average of the accuracies of the four folds gives the final good and bad accuracy. The results are summarized in Table 3.1. The first column in this table denotes the restrictions on foods (if any) regarding the protein and Phe content in the food. The second column denotes the two possible values for K (number of nearest neighbors), which are 4 or 7. The third column denotes the average percentage of foods predicted with a good accuracy (Error $\leq$ 50mg Phe per 100g food) over the four folds. This corresponds to the cumulative sum of the first 10 columns of the histograms shown in Figures 3.5 to 3.22. The fourth column denotes the average percentage of foods predicted with a bad accuracy (Error $>$ 50mg Phe per 100g food) over the four folds. This corresponds to the number of foods populating the last column in the histograms shown in Figures 3.5 to 3.22. So, the sum of the good and bad accuracy in each case must total to a 100%. The good accuracy percentage is determined by taking the average of the good accuracies for each fold. Consider the case of “All Foods” with K=4. The number of accurately predicted foods, with an Error $\leq$ 50mg Phe per 100g food, as seen from Fig. 3.5 are 498, 498, 469 and 508 respectively for the four folds. The total number of foods in each fold are 1269, 1270, 1270 and 1270 respectively. So, the percentage of good accuracies in the four folds are 39.24%, 39.21%, 36.93% and 40.00% respectively. The average value of these accuracies, which is equal to 38.85%, is final value of good accuracy tabulated. Similarly, the bad accuracy is computed to be 61.15% from the number of bad foods in each fold being equal to 771, 772, 801 and 762 respectively.

One could argue that perhaps the reason for the large mistakes in the Phe estimates of certain foods is the lack of similar foods present in the database. To answer this question, we studied the distance of the farthest neighbor considered for estimating Phe of a given test sample. A histogram of these distances are plotted for both the good and bad accuracy foods and overlapped. This allows us to see if there is a relation between distance to the farthest neighbor and accuracy of Phe estimation. Such an analysis helps us explain the density of the database in terms of the accurately and poorly estimated foods. It also helps us determine if the reason for low accuracies of prediction was fewer neighbors closer to the test food.

These distance histograms are shown in Figures 3.23 to 3.31. The x-axis denotes the distance to the farthest neighbor. The y-axis denotes the number of foods with distances to their farthest neighbors lying in the range shown along x-axis. The blue region denotes the foods predicted with good accuracy. The transparent red region denotes the foods predicted with bad accuracy. The purple region denotes overlap of the two categories of foods. As we can see in the Fig. 3.23, there is no clear separation of the distance histograms for the foods predicted with good and bad accuracies. There is a large region of overlap, denoted by purple, which implies that no prediction about the accuracy can be made based solely on the distance of the farthest neighbor. We observe that this is true for all cases of restrictions placed on foods as well (Figures 3.24 to 3.30). This implies that there is no evident relation between the number of close neighbors and the accuracy of Phe estimation. In other words, we can say that the database density is uniform for the foods estimated for Phe with good and bad accuracies. Both the categories of foods have close and far neighbors and the lack of data is not a reason for the low accuracy of prediction for a large fraction of the foods.

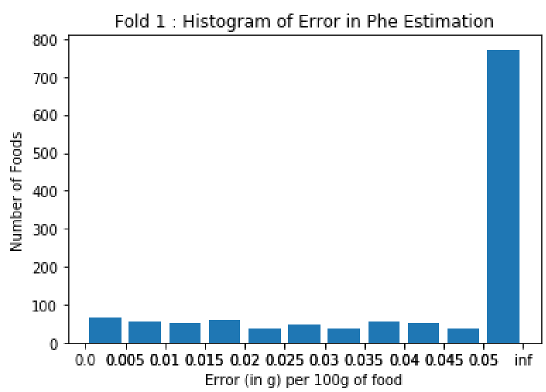
Appendices A and B lists the various foods predicted with good and bad accuracy (K=4) respectively.

As discussed in Section 3.1, for our application, low protein and low Phe foods which are the main constituents of a PKU diet are of particular interest. So we study various cases and combinations of these low protein and low Phe restrictions. Foods with protein $\leq$ 2g and 1g per 100g food are considered. And foods with Phe $\leq$ 4mg and 2mg per 1g of food are considered. The algorithm is tested for different combinations of these restrictions, histograms are observed and accuracies tabulated.

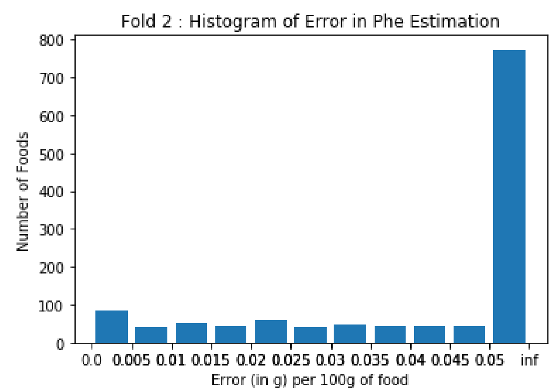
A summary of the experiments is depicted in Figures 3.7 to 3.22 and Table 3.1. The histograms, as seen in Figures 3.7 to 3.22, depict results for the cases with foods restricted to low protein and low Phe cases. By restricting the foods, while we see that the number of total samples reduce, there is also an increase in the fraction of foods estimated with a good accuracy. Let us take a look at the Fig. 3.21. These histograms describe the foods with protein $\leq$ 1g per 100g food and Phe $\leq$ 2mg per 1g of food for the case of K=4. As seen in the figure, the average number of foods per fold is around 140. Equivalently, the total number of foods is 563. The average number of foods estimated with a good accuracy, denoted by the first ten columns of the histogram, is 107 per fold. In other words, on an average over 76% of the 563 foods are estimated with a good accuracy (Error $\leq$ 50mg Phe per 100g food). Also note that on an average, 30 of these foods are estimated with a great accuracy (Error $\leq$ 5mg Phe per 100g food) as seen by the first columns of the histograms for each fold. So, we can see that by restricting the foods to those containing low protein and low Phe, the accuracy of Phe estimation increases. This is important to note because low protein and low Phe foods are the main constituents of a PKU diet. The corresponding results for the case of K=7 as seen in Fig. 3.22 are similar to the K=4 case with a 74% good accuracy of prediction.

These results have also been summarized in Table 3.1. For example, consider the case of "Foods with protein $\leq$1g per 100g of food" with K=4. The number of accurately predicted foods, with an Error $\leq$ 50mg Phe per 100g food, as seen from Fig. 3.17 are 106, 103, 101 and 128 respectively for the four folds. The total number of foods in each fold are 131, 143, 132 and 157 respectively. So, the percentage of good accuracies in the four folds are 80.92%, 72.03%, 76.52% and 81.53% respectively. The average value of these accuracies, which is equal to 77.75%, is the final value of good accuracy tabulated. Similarly, the bad accuracy is computed to be 22.25% from the number of bad foods in each fold being equal to 25, 40, 31 and 29 respectively.

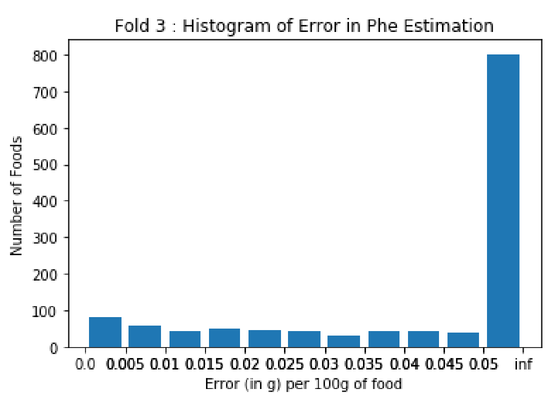
As seen from the Table 3.1, the average percentage of good accuracies increase as one places restrictions on the foods. When all the foods are considered, the good accuracy is only about 38.85%. However, by restricting the protein to 1g per 100g food, we can achieve 77.75% foods predicted with good accuracy (Error $\leq$ 50mg Phe per 100g food). Such restricted foods are of particular interest in our application as they mainly constitute a PKU diet.



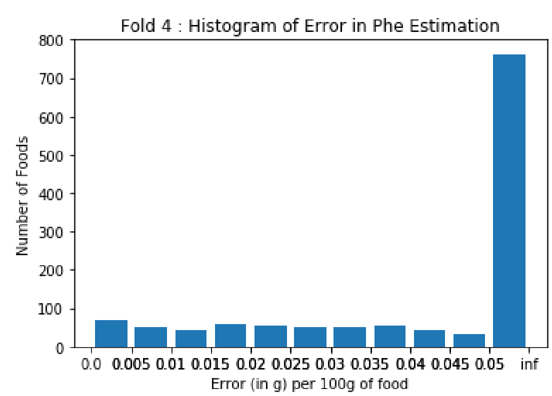
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

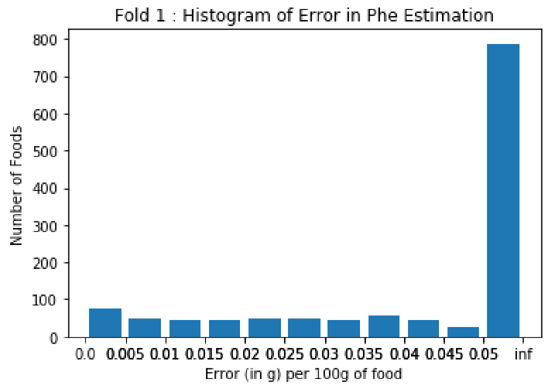


(c) Histogram for Fold 3.

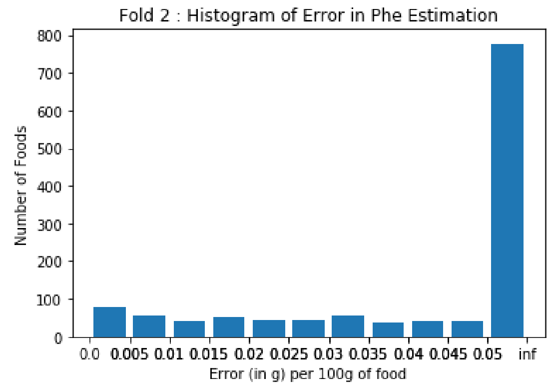


(d) Histogram for Fold 4.

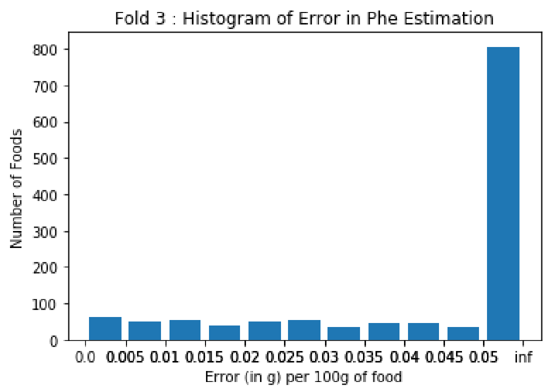
Fig. 3.5. Histograms of Error in Phe Estimation for All Foods (K=4)



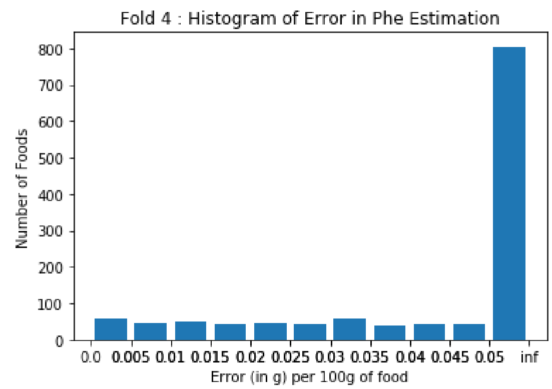
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

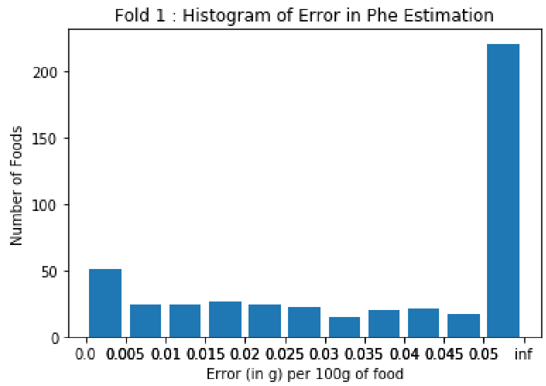


(c) Histogram for Fold 3.

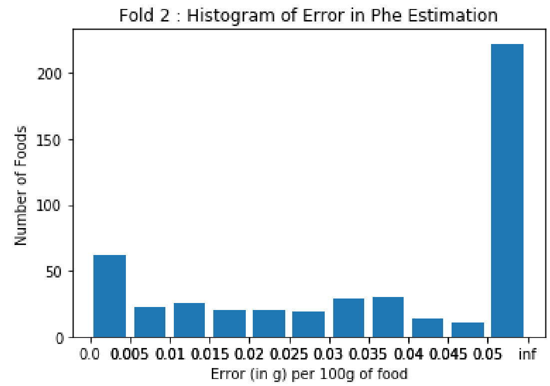


(d) Histogram for Fold 4.

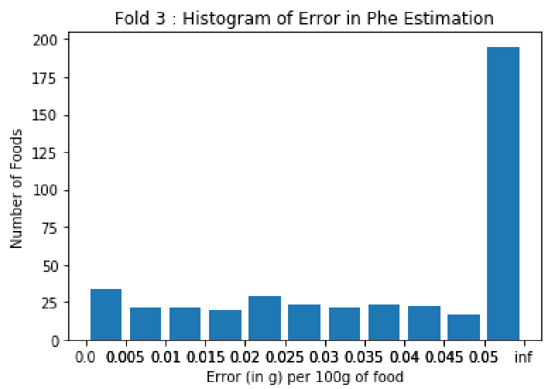
Fig. 3.6. Histograms of Error in Phe Estimation for All Foods (K=7)



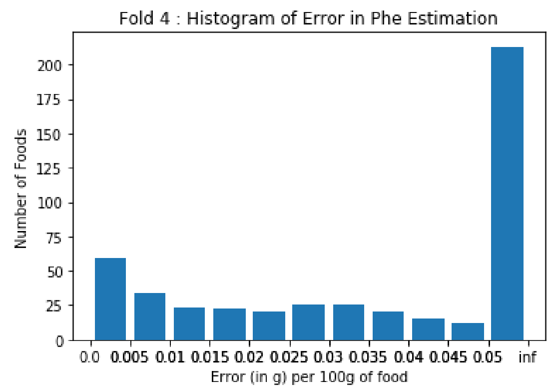
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

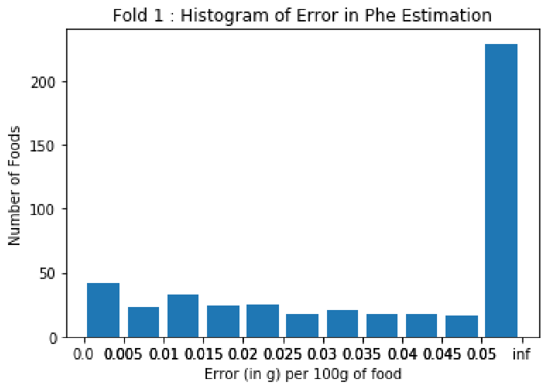


(c) Histogram for Fold 3.

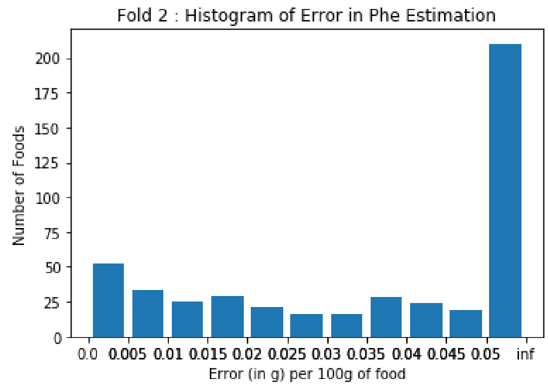


(d) Histogram for Fold 4.

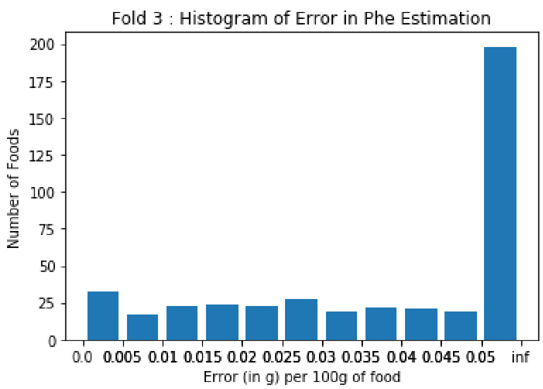
Fig. 3.7. Histograms of Error in Phe Estimation for Foods with Phe $\leq$ 4mg/g (K=4)



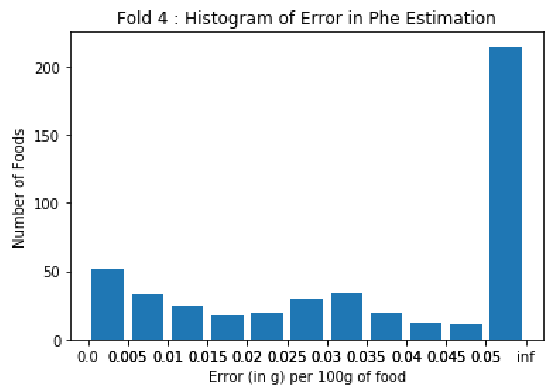
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

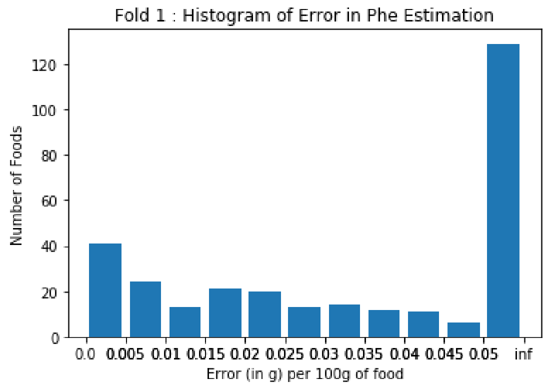


(c) Histogram for Fold 3.

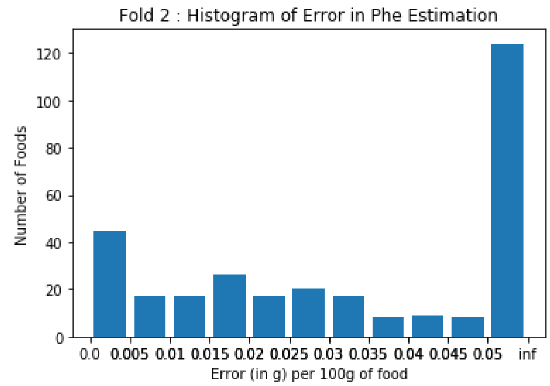


(d) Histogram for Fold 4.

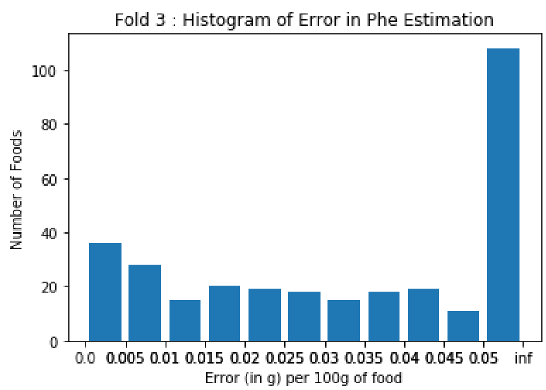
Fig. 3.8. Histograms of Error in Phe Estimation for Foods with Phe $\leq$ 4mg/g (K=7)



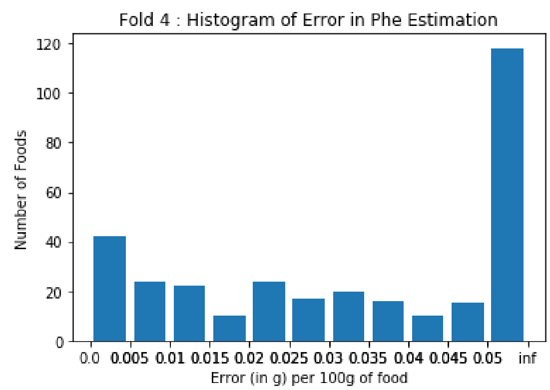
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

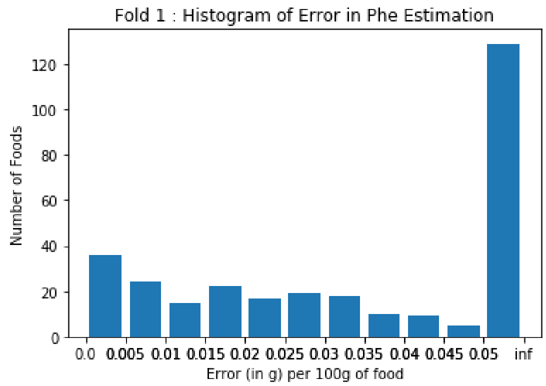


(c) Histogram for Fold 3.

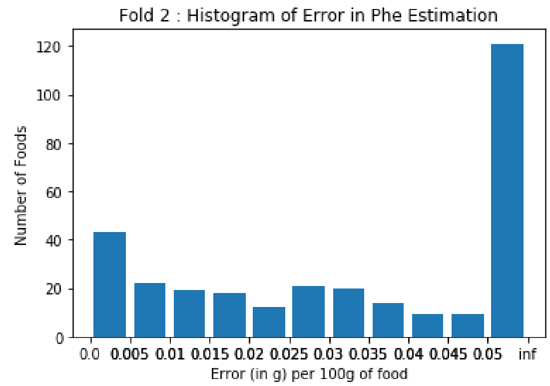


(d) Histogram for Fold 4.

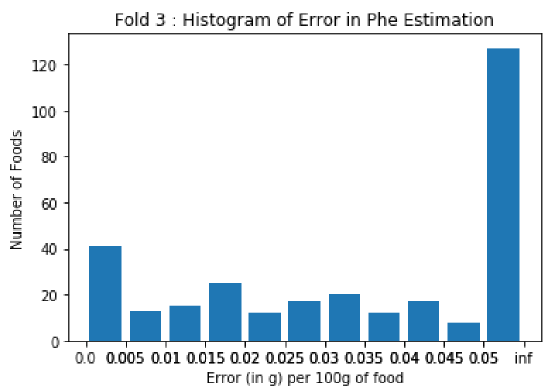
Fig. 3.9. Histograms of Error in Phe Estimation for Foods with Phe $\leq$ 2mg/g (K=4)



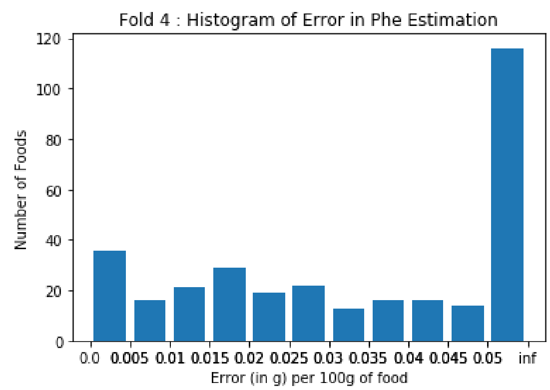
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

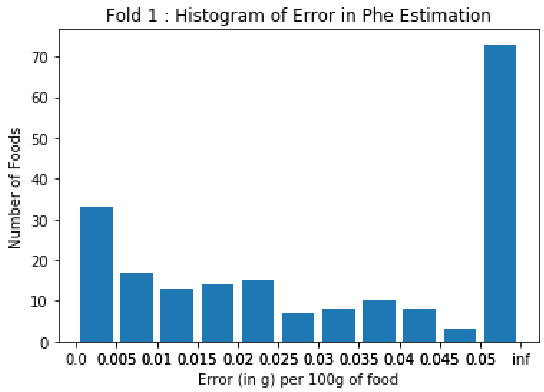


(c) Histogram for Fold 3.

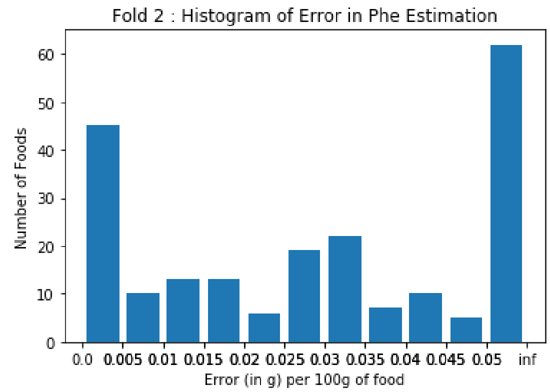


(d) Histogram for Fold 4.

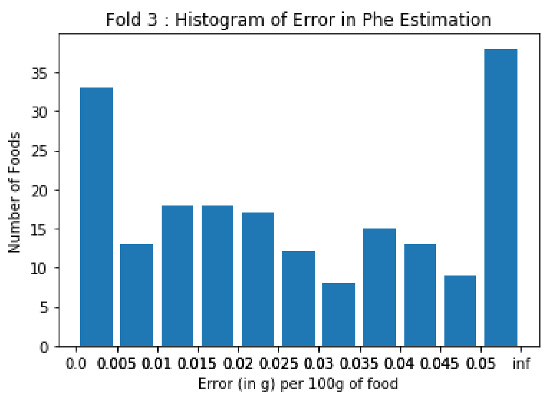
Fig. 3.10. Histograms of Error in Phe Estimation for Foods with Phe $\leq$ 2mg/g (K=7)



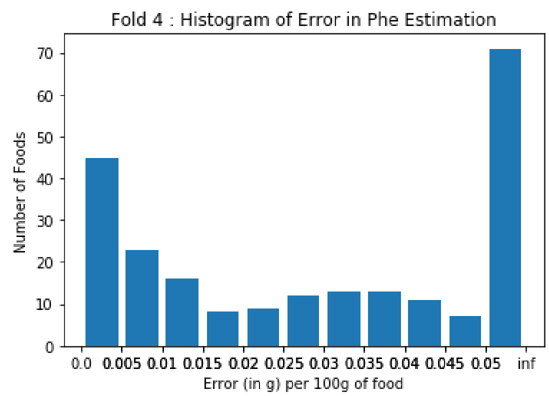
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

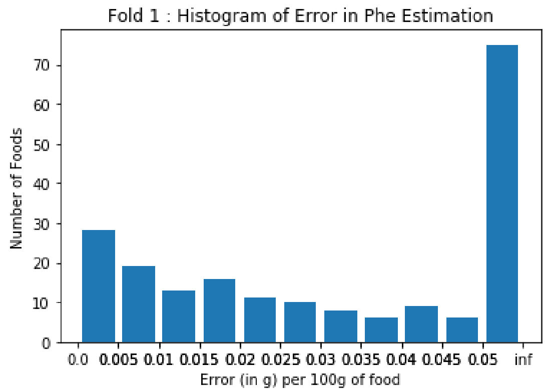


(c) Histogram for Fold 3.

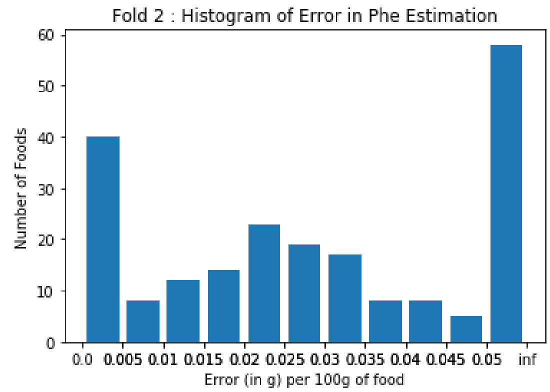


(d) Histogram for Fold 4.

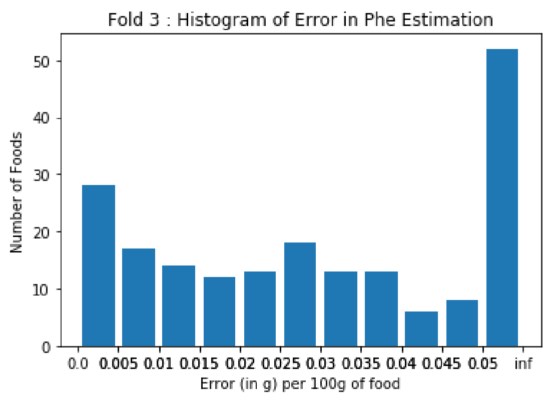
Fig. 3.11. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 2g/100g (K=4)



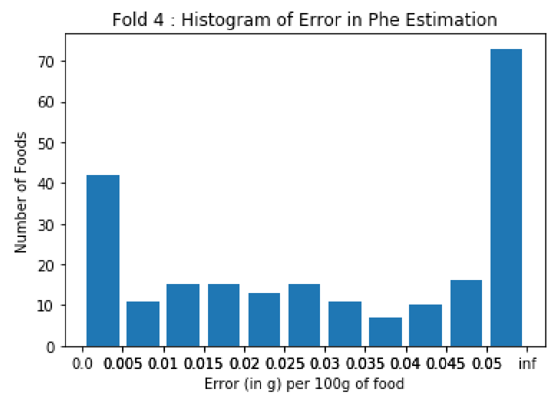
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

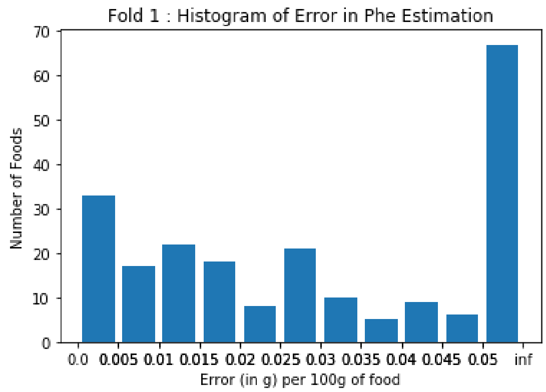


(c) Histogram for Fold 3.

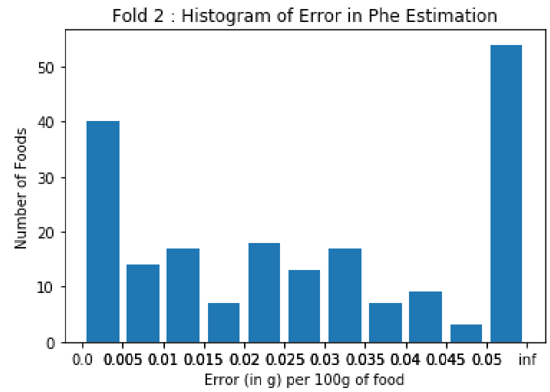


(d) Histogram for Fold 4.

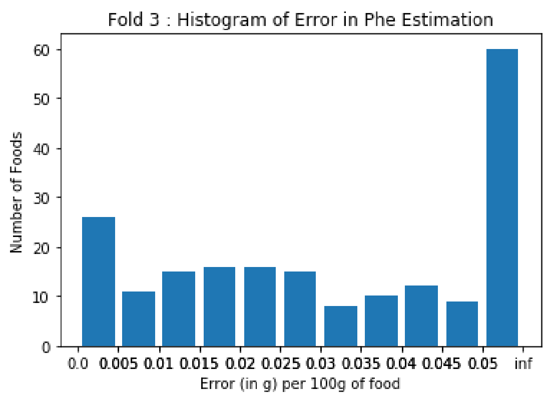
Fig. 3.12. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 2g/100g (K=7)



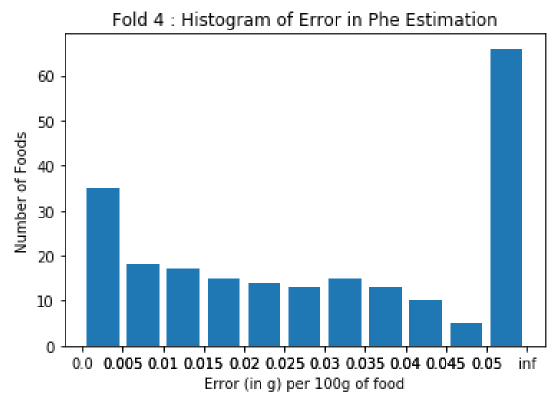
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

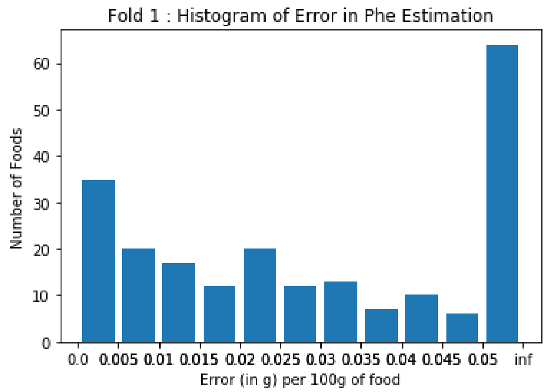


(c) Histogram for Fold 3.

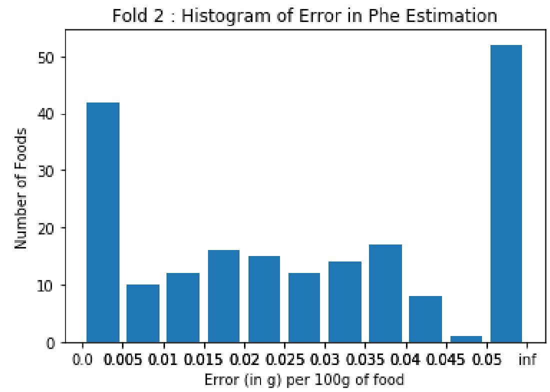


(d) Histogram for Fold 4.

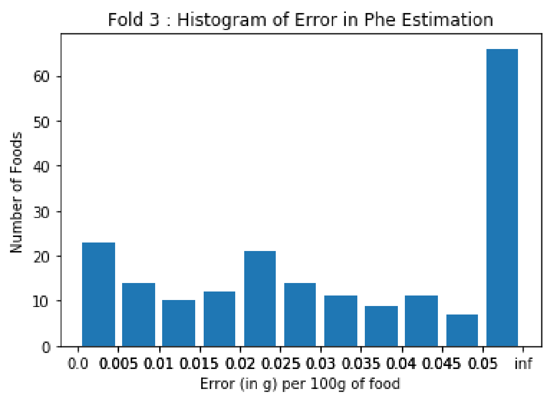
Fig. 3.13. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 2g/100g and Phe $\leq$ 4mg/g (K=4)



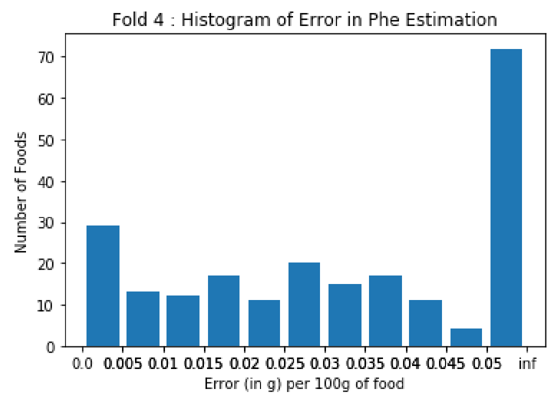
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

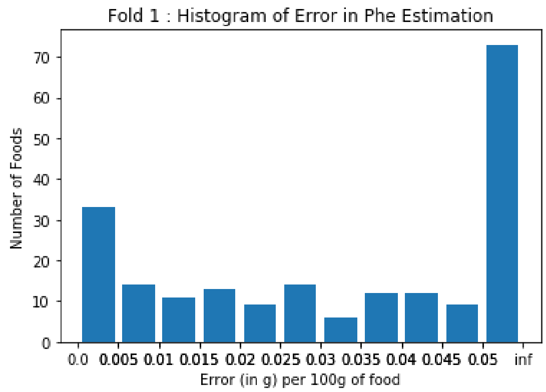


(c) Histogram for Fold 3.

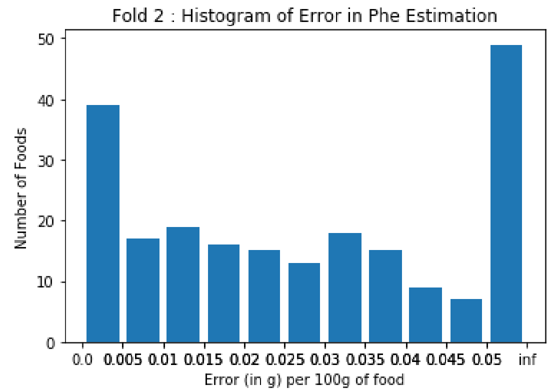


(d) Histogram for Fold 4.

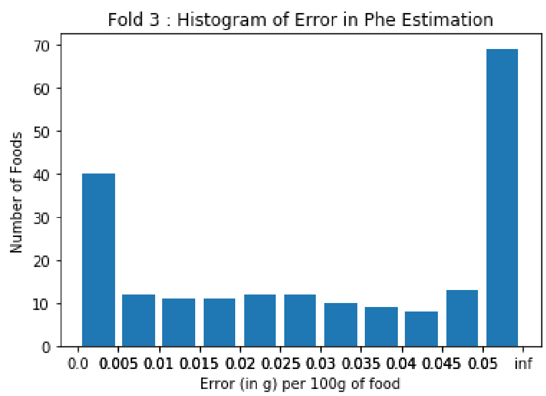
Fig. 3.14. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 2g/100g and Phe $\leq$ 4mg/g (K=7)



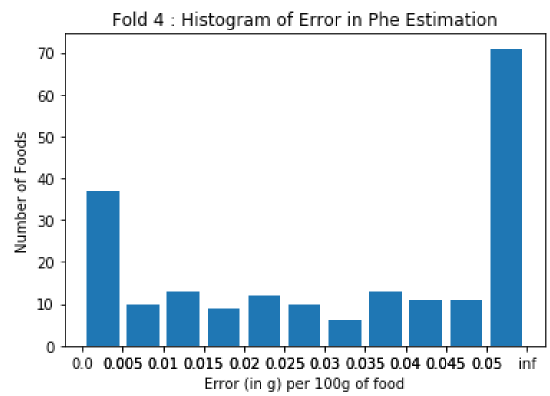
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

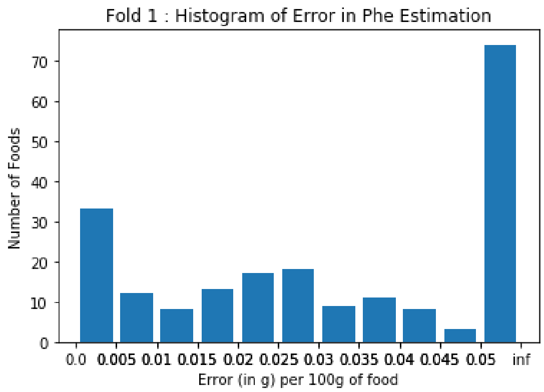


(c) Histogram for Fold 3.

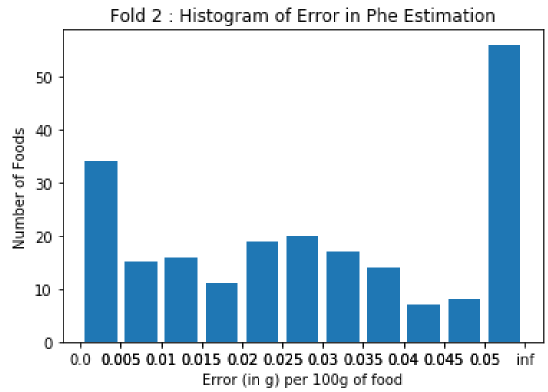


(d) Histogram for Fold 4.

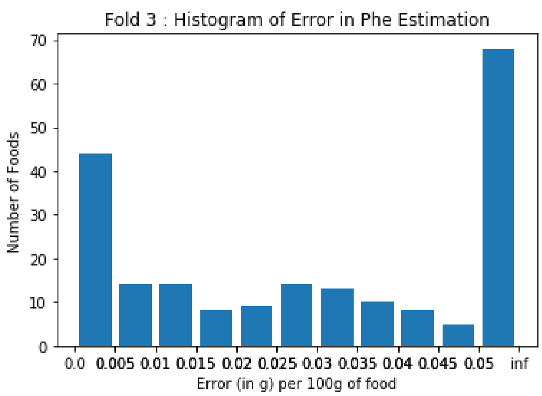
Fig. 3.15. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 2g/100g and Phe $\leq$ 2mg/g (K=4)



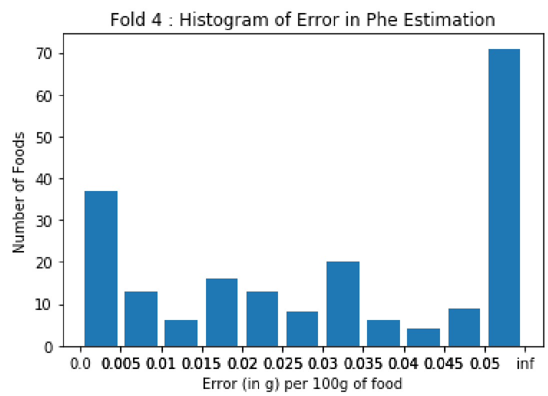
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

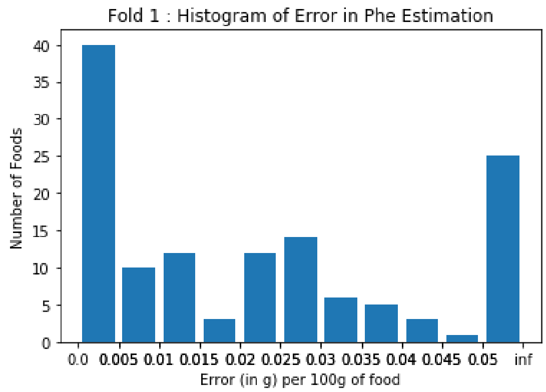


(c) Histogram for Fold 3.

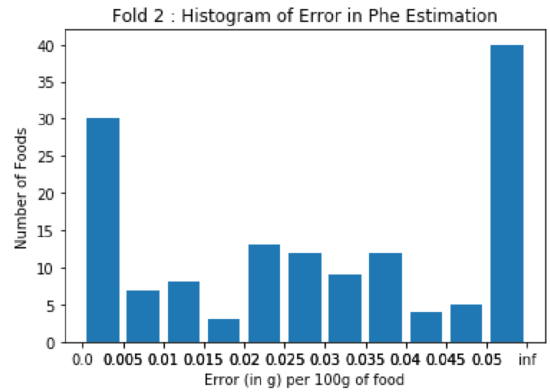


(d) Histogram for Fold 4.

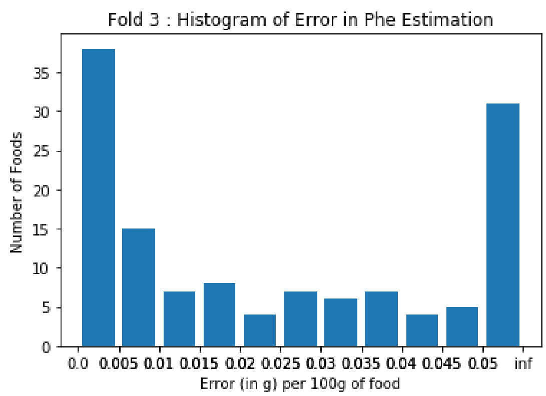
Fig. 3.16. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 2g/100g and Phe $\leq$ 2mg/g (K=7)



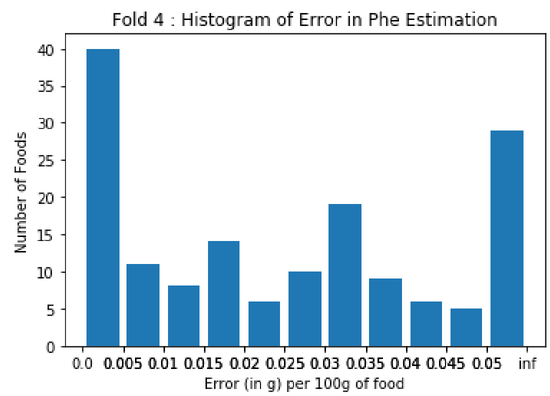
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

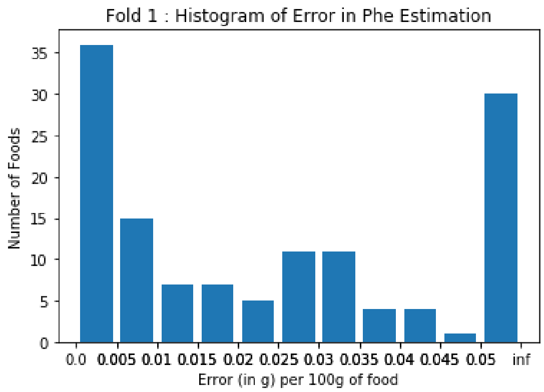


(c) Histogram for Fold 3.

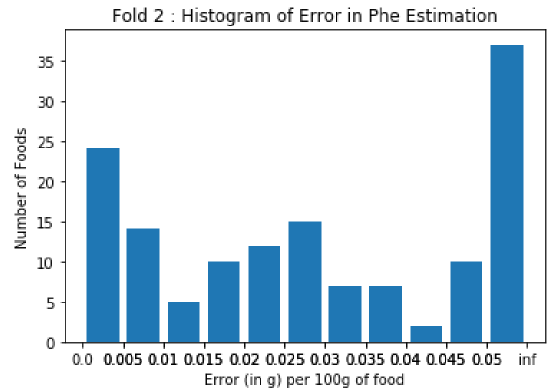


(d) Histogram for Fold 4.

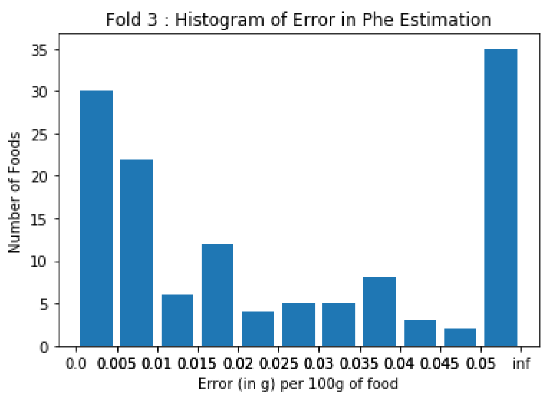
Fig. 3.17. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 1g/100g (K=4)



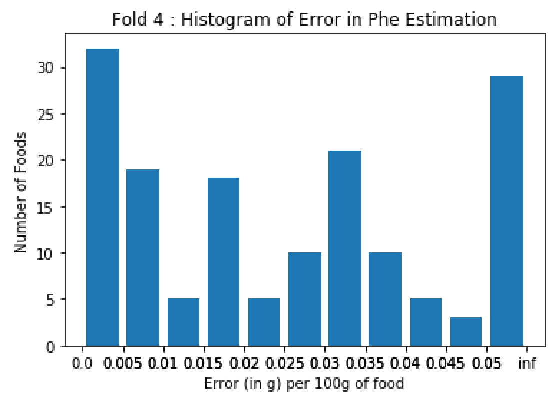
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

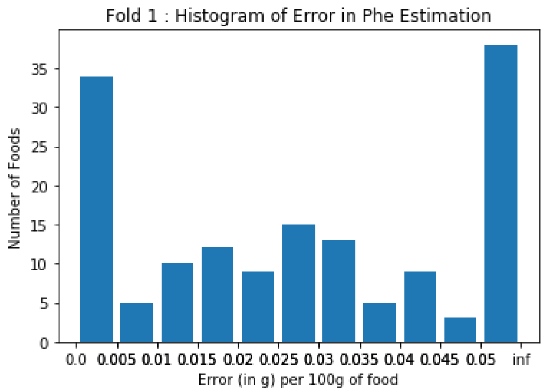


(c) Histogram for Fold 3.

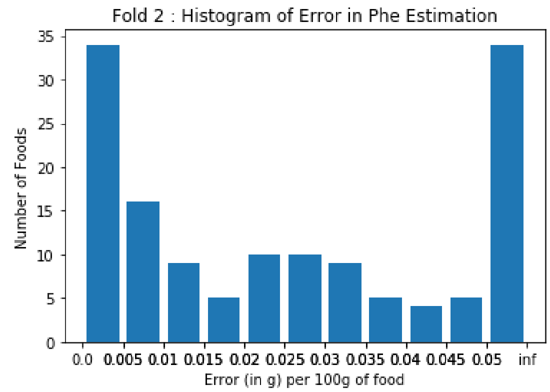


(d) Histogram for Fold 4.

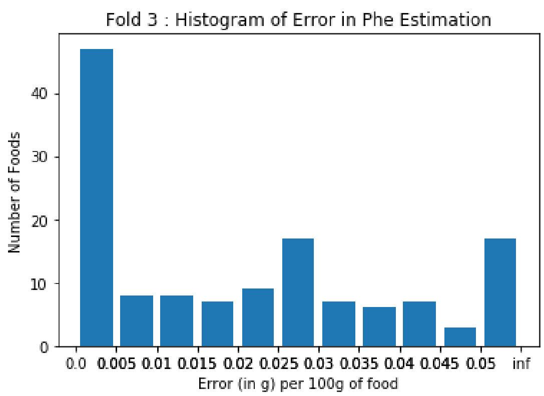
Fig. 3.18. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 1g/100g (K=7)



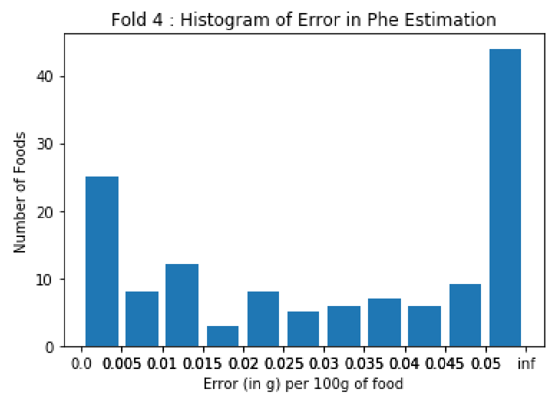
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

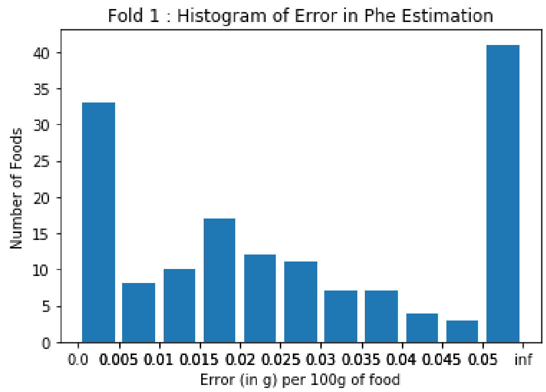


(c) Histogram for Fold 3.

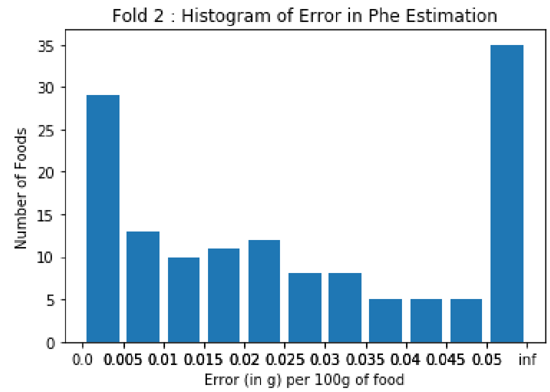


(d) Histogram for Fold 4.

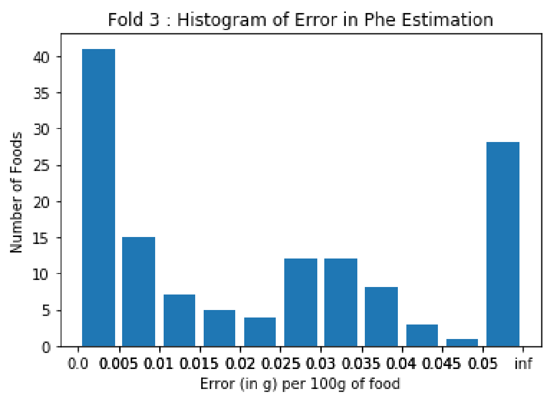
Fig. 3.19. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 1g/100g and Phe $\leq$ 4mg/g (K=4)



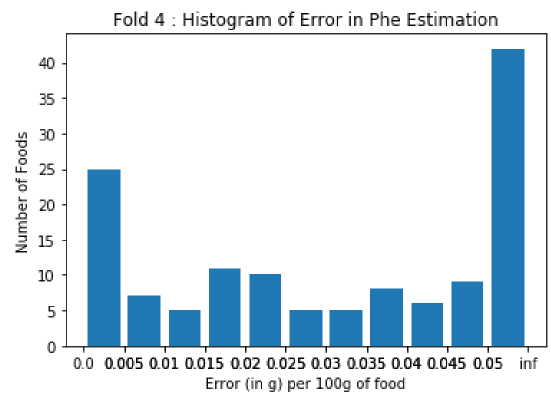
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

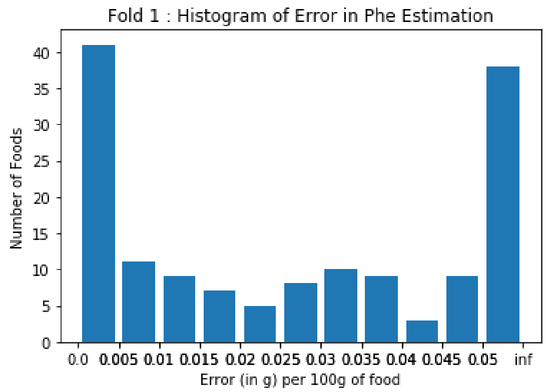


(c) Histogram for Fold 3.

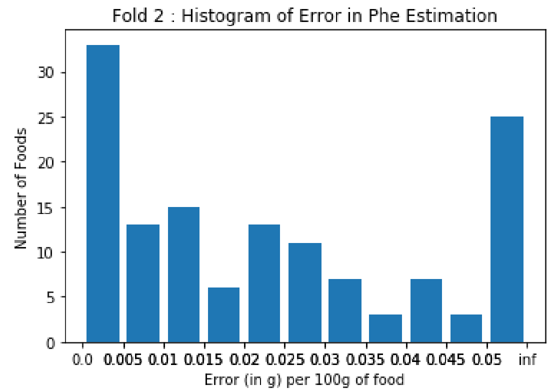


(d) Histogram for Fold 4.

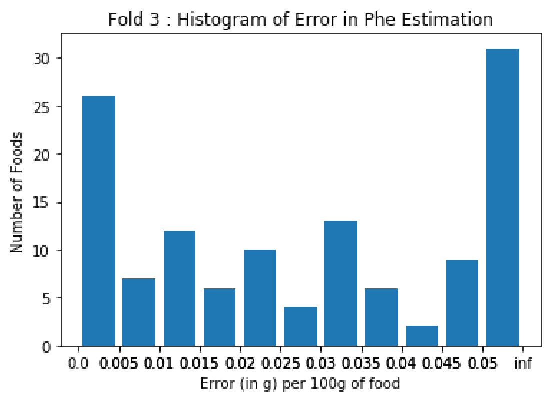
Fig. 3.20. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 1g/100g and Phe $\leq$ 4mg/g (K=7)



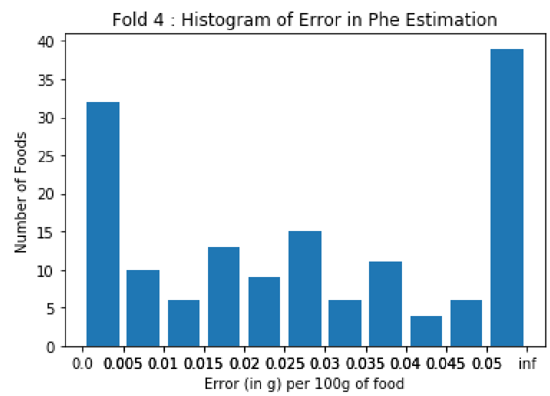
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

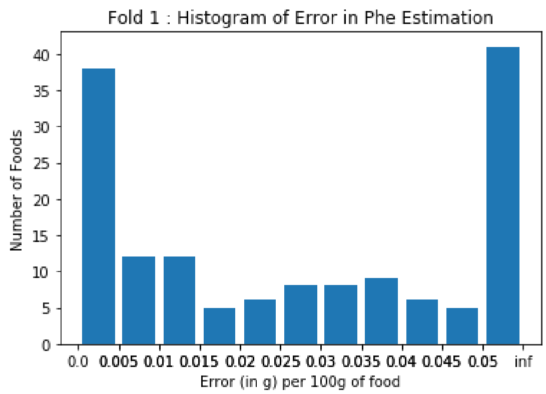


(c) Histogram for Fold 3.

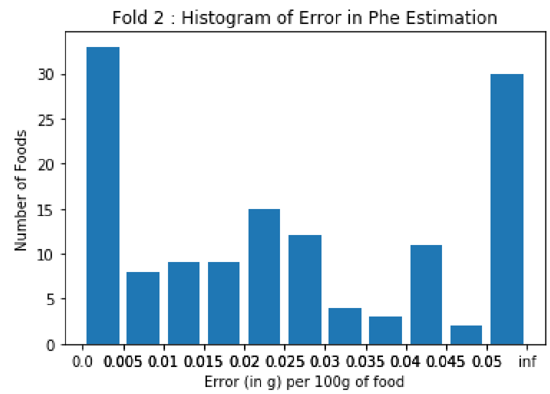


(d) Histogram for Fold 4.

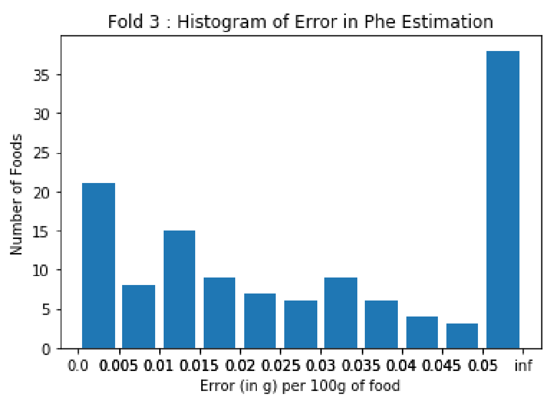
Fig. 3.21. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 1g/100g and Phe $\leq$ 2mg/g (K=4)



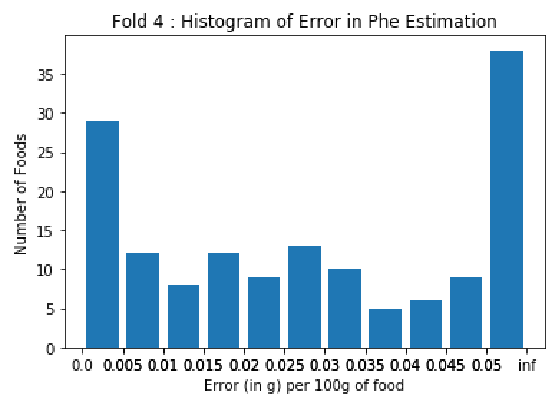
(a) Histogram for Fold 1.



(b) Histogram for Fold 2.

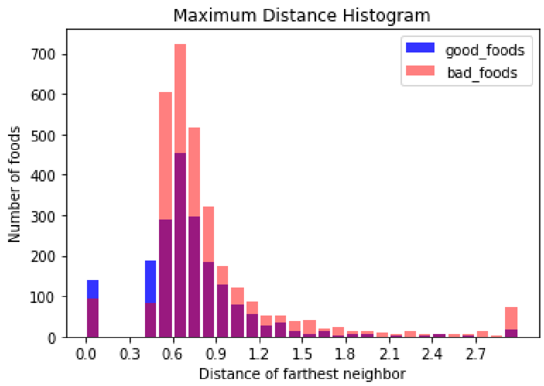


(c) Histogram for Fold 3.

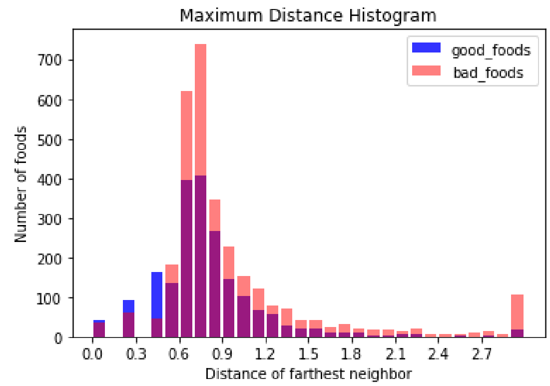


(d) Histogram for Fold 4.

Fig. 3.22. Histograms of Error in Phe Estimation for Foods with Protein $\leq$ 1g/100g and Phe $\leq$ 2mg/g (K=7)

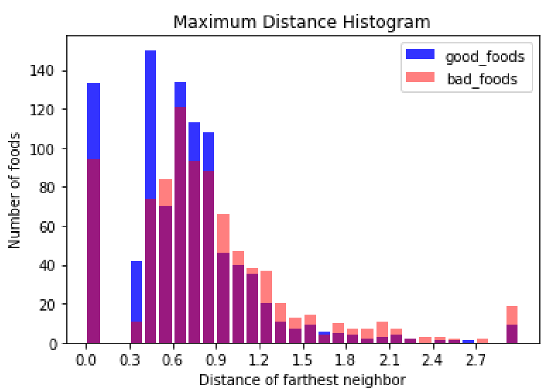


(a) K=4

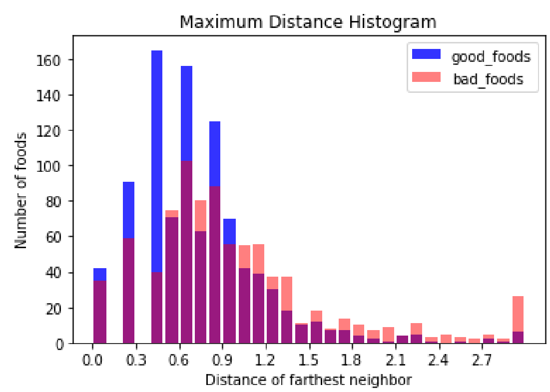


(b) K=7

Fig. 3.23. Histograms of Distance to Farthest Neighbor for All Foods

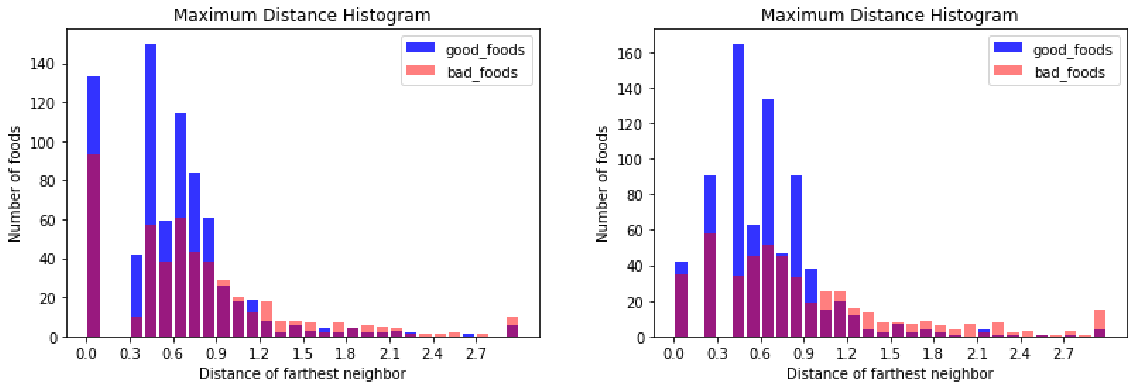


(a) K=4



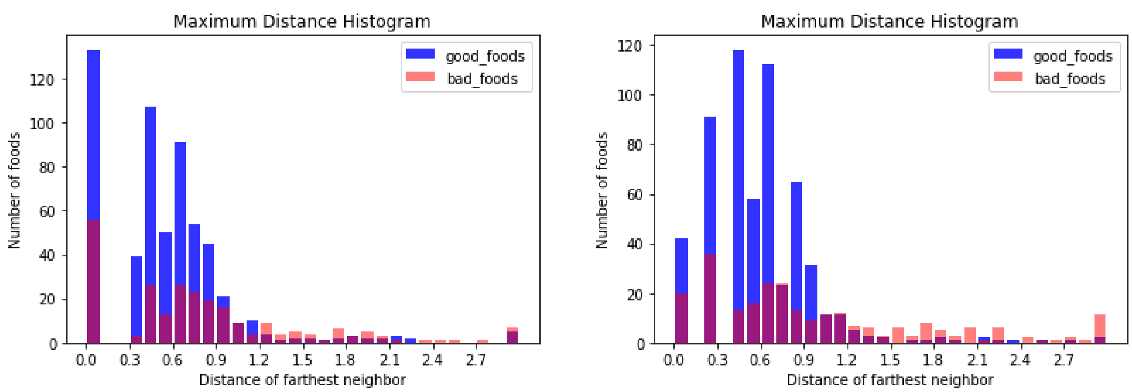
(b) K=7

Fig. 3.24. Histograms of Distance to Farthest Neighbor for Foods with Phe $\leq$ 4mg/g



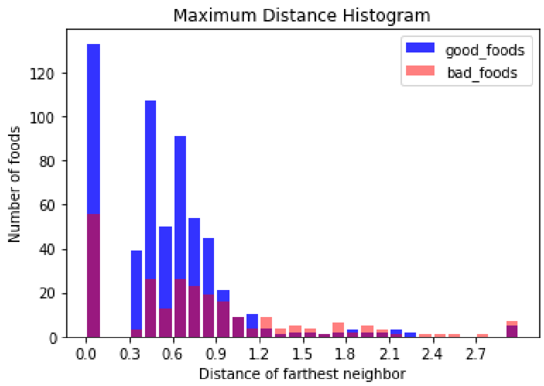
(a) K=4 (b) K=7

Fig. 3.25. Histograms of Distance to Farthest Neighbor for Foods with Phe $\leq$ 2mg/g

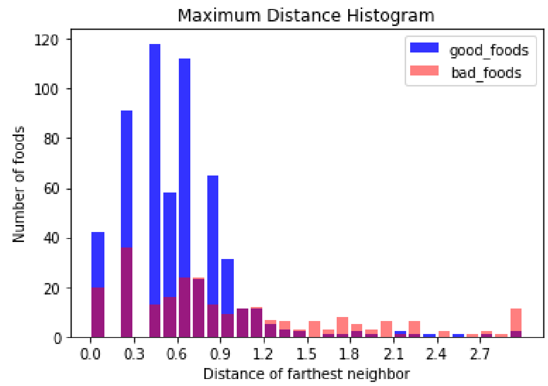


(a) K=4 (b) K=7

Fig. 3.26. Histograms of Distance to Farthest Neighbor for Foods with Protein $\leq$ 2g/100g

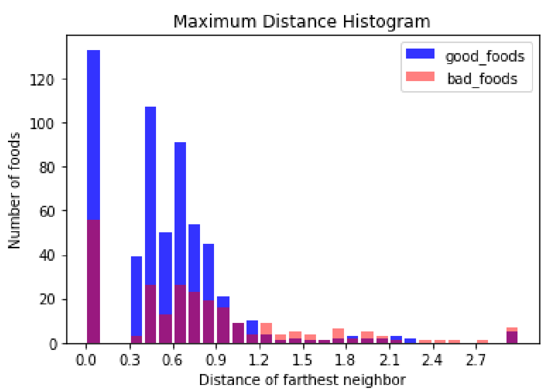


(a) K=4

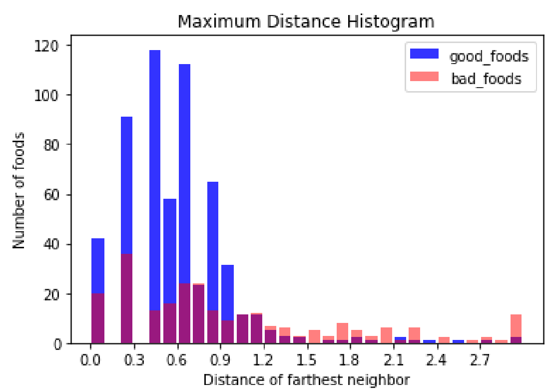


(b) K=7

Fig. 3.27. Histograms of Distance to Farthest Neighbor for Foods with Protein $\leq$ 2g/100g and Phe $\leq$ 4mg/g

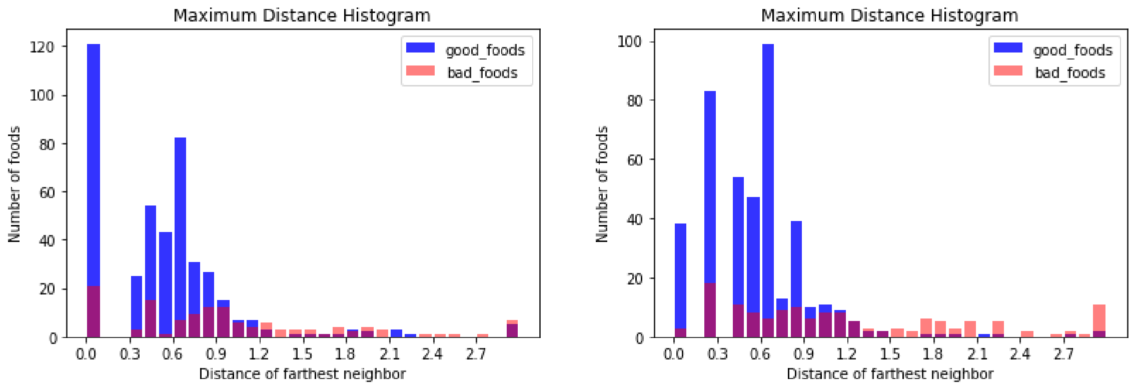


(a) K=4



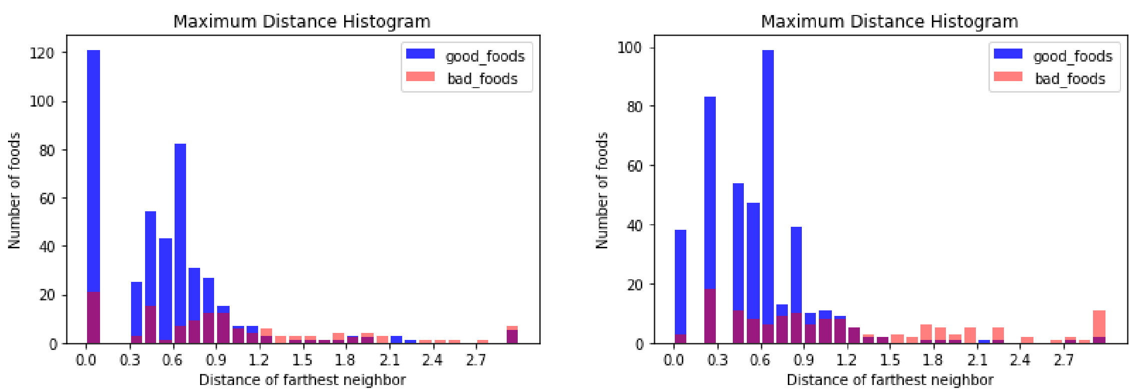
(b) K=7

Fig. 3.28. Histograms of Distance to Farthest Neighbor for Foods with Protein $\leq$ 2g/100g and Phe $\leq$ 2mg/g



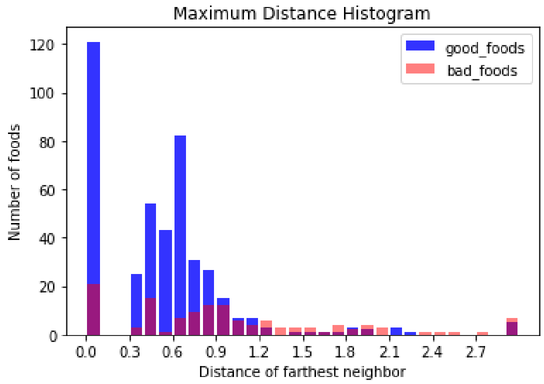
(a) K=4 (b) K=7

Fig. 3.29. Histograms of Distance to Farthest Neighbor for Foods with Protein $\leq$ 1g/100g

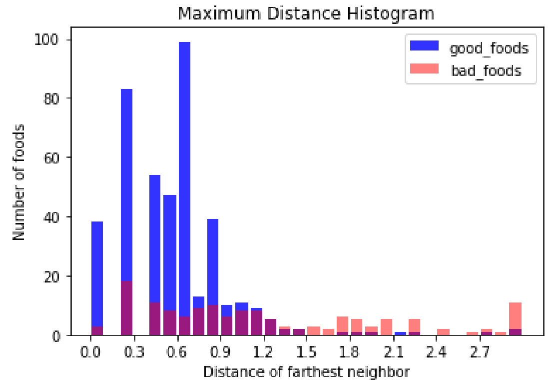


(a) K=4 (b) K=7

Fig. 3.30. Histograms of Distance to Farthest Neighbor for Foods with Protein $\leq$ 1g/100g and Phe $\leq$ 4mg/g



(a) K=4



(b) K=7

Fig. 3.31. Histograms of Distance to Farthest Neighbor for Foods with Protein $\leq$ 1g/100g and Phe $\leq$ 2mg/g

Table 3.1.
Percentage of Foods with Good Accuracies {Error $\leq \pm$50mg per 100g of Food} and Bad Accuracies {Error $> \pm$50mg per 100g of Food } for Machine Learning Approach.

Restrictions on Foods	**K**	**Good Accuracy (%)**	**Bad Accuracy(%)**
All Foods	4	38.85	61.15
	7	37.51	62.49
Foods with Phe $\leq$4mg per g of food	4	53.59	46.41
	7	53.51	46.49
Foods with Phe $\leq$2mg per g of food	4	61.25	38.75
	7	60.11	39.89
Foods with protein $\leq$2g per 100g of food	4	70.93	29.07
	7	69.13	30.87
Foods with protein $\leq$2g per 100g of food and Phe $\leq$4mg per g of food	4	70.42	29.58
	7	69.58	30.42
Foods with protein $\leq$2g per 100g of food and Phe $\leq$2mg per g of food	4	68.42	31.58
	7	67.61	32.39
Foods with protein $\leq$1g per 100g of food	4	77.75	22.25
	7	76.56	23.44
Foods with protein $\leq$1g per 100g of food and Phe $\leq$4mg per g of food	4	76.37	23.63
	7	74.05	25.95
Foods with protein $\leq$1g per 100g of food and Phe $\leq$2mg per g of food	4	76.46	23.54
	7	73.82	26.18

3.6 Conclusion

We have proposed to use the KNN Machine Learning method to estimate the Phe content of a food using its nutrition facts. The dataset used for training is the USDA food nutrient database [7]. We used cross-validation to determine the best values of K as 4 or 7.

In the general case of all foods, the Machine Learning approach has an accuracy of about 38%. In other words, only 38% of the total 5079 foods were estimated with a good accuracy (Error $\leq$ 50mg Phe per 100g food). However, these foods include high protein foods like meats, dairy products, nuts and aspartame which are strictly prohibited in a PKU diet. By restricting the foods to those with low protein and low Phe (foods typically consumed by a PKU patient), we achieve higher food percentages with a good accuracy (77% foods with protein $\leq$ 1g and Phe $\leq$ 2mg with error $\leq$ 50mg Phe). These are the foods that matter.

By looking at the distance histograms in Figures 3.23 to 3.31, we see that there is not much distinction between the histograms for the foods estimated with good and bad accuracies. The distribution is quite similar for both categories. In other words, both categories have the farthest neighbors at small and large distances. So we can say that the data is not sparse or that the database is not unevenly distributed for the good and bad accuracy cases. This distance to neighbors is not related to the accuracy of estimation. Hence, we can conclude that the low accuracy of estimation using ML approach is not because database is incomplete but because information is actually incomplete in nutrition facts alone. Thus, we might be able to improve the accuracy of Phe Estimation only by incorporating additional information (e.g., ingredients list).

4. COMPARISON OF THE PROPOSED TWO METHODS AND CONCLUSION

In the following discussion, we compare the numerical results obtained from the ML approach (Chapter 3) with respect to those numerical results obtained from the Mathematical approach (Chapter 2). For this, the 20 commercial foods [20] studied in the Mathematical approach are used. The ML Phe estimate is calculated by K-NN classification (K=7). We used the rounded nutrient values from the food labels and performed an exhaustive search for the nearest neighbors in the entire USDA database. Since 6 of the 20 foods considered were themselves present in the database, they were removed prior to the search.

Table 4.1 shows the actual Phe values per serving for the 20 foods from either USDA or the Low-protein food database. It contains the phe estimates obtained from the Mathematical method and the difference between this estimate and the actual Phe (that gives the lowest difference). All values of Phe are in mg per serving of the food. We can see that the results obtained from Mathematical method are fairly accurate. The cells shaded yellow indicate actual Phe values of those foods which lie within the final interval predicted by the Mathematical method. The error between the prediction and actual Phe as seen from Table 4.1 is less than $\pm 20mg$ per serving for 16 out of the 20 foods. The maximum error is about -32mg (Food # 8) or 33mg (Food # 3). This indicates a good accuracy for about 80% of the foods by the mathematical approach. So we can say that this approach has a good accuracy for Phe estimation.

Similarly, Table 4.2 shows the actual Phe values, ML Phe estimates and the difference between the estimate and actual Phe value. We see that the error is less than $\pm 20mg$ for around 13 foods only. This comprises about 65% of the total cases. Also, the errors for the ML approach can be as large as -126.78mg (Food # 19) or

-106.77mg (Food # 13). So we may say that the errors in the ML approach are larger than those seen in the Mathematical method.

We also computed the difference in Phe estimates (in mg) obtained from the two approaches for a serving of the foods. Since Mathematical method is more accurate, we also checked if the ML Phe estimate lies in the Phe interval resulting from the Mathematical method (Table 2.2). The results are summarized in Tables 4.3.

Table 4.3 shows the results of comparison between the two methods for the case when the ML approach estimates Phe (in mg) per 100g of the food. The first column in the table denotes the food number as referenced in Table 2.1. The second column denotes the Phe content estimated with ML approach (rounded to two decimal places). The cells shaded yellow indicate those foods whose ML Phe estimates lie within the Phe interval predicted by Mathematical method. The third column denotes the difference between Phe estimates of ML and Mathematical methods (in mg).

As seen from the Tables 4.3 the number of foods whose ML Phe estimate conforms to the Mathematical Phe range estimate are only about 25% of the total foods. In some cases, the two approaches agree very well with each other with a difference in estimates of only -0.3mg (for Food # 11 in Table 4.3). Also, around 50% of the food estimates are well within a difference of ±10mg per serving. However, for most of the cases (over 75% of the foods), we see that the Phe estimates calculated with the ML approach do not agree with the predicted Phe intervals from the Mathematical method. Also, the difference between the two estimates can be as high as -131.21mg (for Food # 19 in Table 4.3). We conclude that the 3-step Mathematical method is very accurate. In fact, combining the three steps gave us improved accuracies than those obtained by the individual methods proposed earlier.

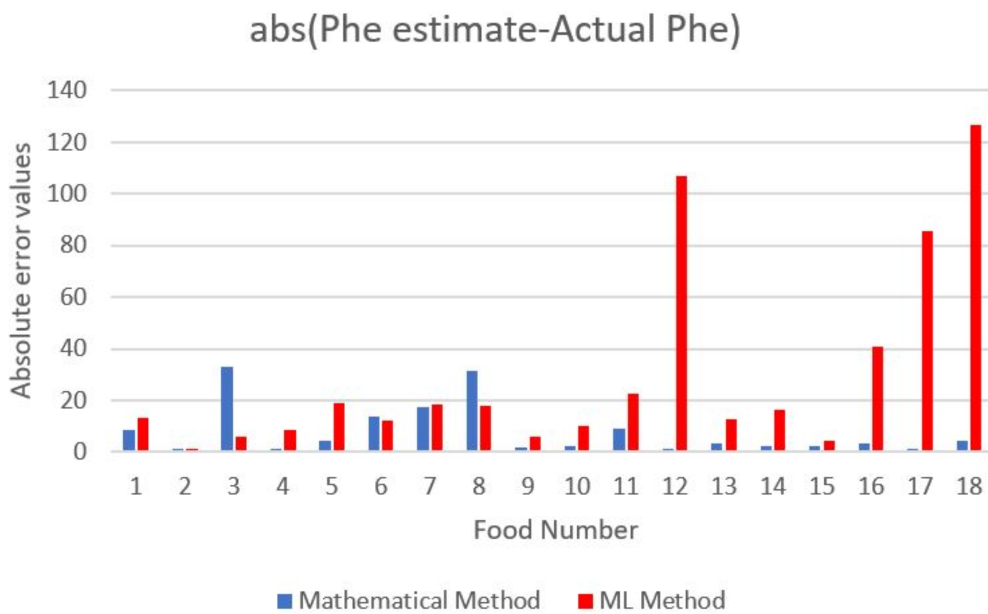


Fig. 4.1. Difference Between Ground Truth and Phe Estimates Obtained by Mathematical and ML Method.

To better understand the nature of these errors we plot the difference between the actual Phe estimates and those obtained by Mathematical and ML method as shown in Fig. 4.1. We can see that in many cases, the error obtained by ML approach is comparable to the error from the Mathematical method. It may even be smaller than the Mathematical error in some cases. However, an important observation is in quite a few cases, the prediction can be terribly wrong as seen for Foods # 12, 17 and 18 from Fig. 4.1. This unpredictability is rendered in ML method because Phe estimates are predicted without any error bounds like those in the Mathematical method. In order to combat this, one may choose to use the results of ML method along with the results of Step 1 of the Mathematical method. Ascertaining that the ML Phe estimate lies in Step 1 interval would add more reliability to the process.

We see that the ML approach is not very accurate in the general case of all foods. However, it is important to realize that this set contains a lot of high protein foods which are strictly prohibited in a PKU diet and hence, is not very relevant to our study. On the other hand, the ML method performs fairly accurately in the foods with protein and Phe restricted to low values. As we saw from Chapter 3, ML method performs with an accuracy up to 77% for the low-protein, low-Phe foods which are the relevant foods for our target disorder.

Additionally, the ease of use and simplicity of the ML approach makes it very desirable for user applications. It eliminates the need to enter the ingredients in order of their increasing weights as required by the Mathematical method. In fact, the need to enter data about ingredients, nutrients and serving size makes the third step of the Mathematical method cumbersome. A search for the ingredients in the database is often laborious, ambiguous and may even sometimes be futile. In such scenarios, the ML approach serves as a feasible alternative apart from the Step 1 of the Mathematical method which only yields a crude range for Phe. However, the ML method does not give any error bounds on the estimate. So, there can be significant risks to using this method.

One might argue that the ML approach may suffer from the lack of data or in other words, the small size of the database considered (5079 foods) [7]. It could also be said that the errors could be attributed to the lack of close neighbors. However, an analysis of the Figures 3.23 to 3.31 shows that this is not the case. Thus, we believe that the problem is ill-posed and the ML approach for Phe estimation cannot be further improved without using additional data (like the ingredients list). However, it might still be worthwhile to attempt this problem using Deep Neural Networks (DNNs). Ideally, we should test the results on a completely new database, so that the test results are not biased by the training data.

Table 4.1.
Comparison of the Mathematical Approach with Ground Truth.

#	**USDA database Actual Phe (in mg)**	**Low-protein database Actual Phe (in mg)**	**Mathematical Phe estimate (in mg)**	**Mathematical Phe estimate-Actual Phe**
1	81.60	75	66.65	-8.35
2	4.42	10.2	3.89	-0.53
3	113.4	131.86	165.11	33.25
4	68.32	66.90	68.07	-0.25
5	175.84	165	180.19	4.35
6	116.82	107	93.17	-13.83
7	N/A	6	23.17	17.17
8	N/A	238	206.31	-31.69
9	N/A	1.93	3.96	2.03
10	N/A	11	13.07	2.07
11	N/A	N/A	150.87	-
12	N/A	6	14.86	8.86
13	N/A	120	118.67	-1.33
14	N/A	23.76	20.23	-3.53
15	N/A	76	78.20	2.20
16	N/A	8	10.14	2.14
17	N/A	5.42	2.24	-3.18
18	N/A	3	2.77	-0.23
19	N/A	284.67	289.10	4.43
20	N/A	N/A	2.33	-

Table 4.2.
Comparison of the Machine Learning Approach with Ground Truth.

#	**USDA database Actual Phe (in mg)**	**Low-protein database Actual Phe (in mg)**	**ML Phe estimate (in mg)**	**ML Phe estimate-Actual Phe**
1	81.60	75	61.57	-13.43
2	4.42	10.2	10.58	0.38
3	113.4	131.86	119.23	5.83
4	68.32	66.90	58.17	-8.73
5	175.84	165	145.86	-19.14
6	116.82	107	128.92	12.1
7	N/A	6	24.19	18.19
8	N/A	238	220.05	-17.95
9	N/A	1.93	8.13	6.2
10	N/A	11	20.88	9.88
11	N/A	N/A	150.57	-
12	N/A	6	28.70	22.70
13	N/A	120	13.23	-106.77
14	N/A	23.76	11.20	-12.56
15	N/A	76	59.61	-16.39
16	N/A	8	12.25	4.25
17	N/A	5.42	46.37	40.95
18	N/A	3	88.58	85.58
19	N/A	284.67	157.89	-126.78
20	N/A	N/A	7.00	-

Table 4.3.
Comparison of the Machine Learning Approach and the Mathematical Approach.

#	Phe Estimated with ML Approach (in mg) per serving	ML Phe Estimate-Mathematical Phe Estimate (in mg) per serving
1	61.57[1]	-5.08
2	10.58	6.69
3	119.23	-45.88
4	58.17[1]	-9.9
5	145.86	-34.33
6	128.92	35.75
7	24.19[1]	1.02
8	220.05	13.74
9	8.13	4.17
10	20.88	7.81
11	150.57[1]	-0.3
12	28.70	13.84
13	13.23	-105.44
14	11.20[1]	-9.03
15	59.61	-18.59
16	12.25	2.11
17	46.37	44.13
18	88.58	85.81
19	157.89	-131.21
20	7.00	4.67

[1]Cells shaded yellow indicate those foods whose ML Phe estimates lie within the Phe interval predicted by Mathematical method.

4.1 Future Work

Optical Character Recognition (OCR) and other computer vision techniques can be employed to make the web applications more user friendly. For example, it could enable the user to click a picture of the food label and the code would be able to extract all the information regarding ingredients, nutrients and serving size. Alternatively, it could read the barcode and access the same information from a database. The accuracy of the Machine Learning method could also be improved. For example, information regarding ingredients could be included in the KNN approach to improve accuracy. Also, different ML techniques like Deep Neural Networks (DNNs) or Regression Analysis can be employed to solve the problem.

REFERENCES

REFERENCES

[1] J. Kim and M. Boutin, “New multipliers for estimating the phenylalanine content of foods from the protein content,” *Journal of Food Composition and Analysis*, vol. 42, pp. 117–119, 2015.

[2] ——, “A list of phenylalanine to protein ratios for common foods,” ECE Technical Reports. Paper 456. Available at: http://docs.lib.purdue.edu/ecetr/456, 2014.

[3] ——, “An approximate inverse recipe method with application to automatic food analysis,” in *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on.* IEEE, 2014, pp. 32–39.

[4] ——, “Estimating the nutrient content of commercial foods from their label using numerical optimization,” in *New Trends in Image Analysis and Processing–ICIAP 2015 Workshops.* Springer, 2015, pp. 309–316.

[5] J. Kim, “Mathematical approaches to food nutrient content estimation with a focus on phenylalanine,” Ph.D. dissertation, Purdue University, 2015.

[6] J. Kim, A. Talikoti, and M. Boutin, “A 3-step process to estimate phenylalanine in commercial foods for pku management,” *IEEE Access*, vol. 6, pp. 30 758–30 765, December 2018.

[7] U.S. Department of Agriculture (USDA), Agricultural Research Service, Nutrient Data Laboratory., “USDA national nutrient database for standard reference, legacy,” Internet: http://www.ars.usda.gov/nutrientdata, April 2018.

[8] J. Campbell and J. Porter, “Dietary mobile apps and their effect on nutritional indicators in chronic renal disease: a systematic review,” *Nephrology*, vol. 20, no. 10, pp. 744–751, 2015.

[9] M. Hingle and H. Patrick, “There are thousands of apps for that: navigating mobile technology for nutrition education and behavior,” *Journal of nutrition education and behavior*, vol. 48, no. 3, pp. 213–218, 2016.

[10] T. Prioleau, E. Moore II, and M. Ghovanloo, “Unobtrusive and wearable systems for automatic dietary monitoring,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2075–2089, 2017.

[11] K. M. Azar, L. I. Lesser, B. Y. Laing, J. Stephens, M. S. Aurora, L. E. Burke, and L. P. Palaniappan, “Mobile applications for weight management: theory-based content analysis,” *American journal of preventive medicine*, vol. 45, no. 5, pp. 583–589, 2013.

[12] J. Chen, J. E. Cade, and M. Allman-Farinelli, "The most popular smartphone apps for weight loss: a quality assessment," *JMIR mHealth and uHealth*, vol. 3, no. 4, 2015.

[13] M. El-Dosuky, M. Rashad, T. Hamza, and A. El-Bassiouny, "Food recommendation using ontology and heuristics," in *International conference on advanced machine learning technologies and applications.* Springer, 2012, pp. 423–429.

[14] G. Agapito, B. Calabrese, I. Care, D. Falcone, P. H. Guzzi, N. Ielpo, T. Lamprinoudi, M. Milano, M. Simeoni, and M. Cannataro, "Profiling basic health information of tourists: Towards a recommendation system for the adaptive delivery of medical certified nutrition contents," in *High Performance Computing & Simulation (HPCS), 2014 International Conference on.* IEEE, 2014, pp. 616–620.

[15] G. Agapito, B. Calabrese, P. H. Guzzi, M. Cannataro, M. Simeoni, I. Car, T. Lamprinoudi, G. Fuiano, and A. Pujia, "Dietos: A recommender system for adaptive diet monitoring and personalized food suggestion," in *2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Oct 2016, pp. 1–8.

[16] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE journal of selected topics in signal processing*, vol. 4, no. 4, pp. 756–766, 2010.

[17] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1261–1271, 2014.

[18] F. Zhu, M. Bosch, N. Khanna, C. J. Boushey, and E. J. Delp, "Multiple hypotheses image segmentation and classification with application to dietary assessment," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 377–388, 2015.

[19] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: a new dataset, experiments, and results," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 588–598, 2017.

[20] A. Talikoti and M. Boutin, "20 commercial foods to test 3-step phenylalanine estimation process," Mar 2019. [Online]. Available: https://purr.purdue.edu/publications/2940/1

[21] D. Stansbury, "The statistical whitening transform, the clever machine," Internet:https://theclevermachine.wordpress.com/2013/03/30/the-statistical-whitening-transform/, March 2013.

[22] A. Talikoti and M. Boutin, "Python code for estimating phenylalanine (phe) using the k-nearest neighbor method." [Online]. Available: https://purr.purdue.edu/publications/3182/1

APPENDICES

A. LIST OF FOODS ESTIMATED WITH GOOD ACCURACY (ERROR $\leq$ 50MG PHE PER 100G OF FOOD) (K=4)

1. Butter, whipped, with salt
2. Butter oil, anhydrous
3. Dessert topping, semi solid, frozen
4. Milk, human, mature, fluid
5. Whey, acid, fluid
6. Whey, sweet, fluid
7. Butter, without salt
8. Cream substitute, flavored, liquid
9. Vinegar, cider
10. Babyfood, GERBER, 2nd Foods, apple, carrot and squash, organic
11. Babyfood, tropical fruit medley
12. Babyfood, vegetables, green beans, junior
13. Babyfood, vegetables, beets, strained
14. Babyfood, vegetables, carrots, strained
15. Babyfood, vegetables, carrots, junior
16. Babyfood, vegetables, sweet potatoes strained
17. Babyfood, vegetables, sweet potatoes, junior
18. Babyfood, vegetables, corn, creamed, strained
19. Babyfood, vegetables, corn, creamed, junior
20. Babyfood, cereal, mixed, with applesauce and bananas, strained
21. Babyfood, cereal, mixed, with applesauce and bananas, junior, fortified
22. Babyfood, cereal, rice, with applesauce and bananas, strained
23. Babyfood, dessert, dutch apple, junior

24. Babyfood, dessert, fruit pudding, orange, strained
25. Babyfood, vegetables, mix vegetables strained
26. Babyfood, beverage, GERBER, GRADUATES, FRUIT SPLASHERS
27. Babyfood, corn and sweet potatoes, strained
28. Babyfood, fruit, banana and strawberry, junior
29. Babyfood, banana with mixed berries, strained
30. Fat, beef tallow
31. Lard
32. Salad dressing, mayonnaise type, regular, with salt
33. Salad dressing, french dressing, reduced fat
34. Salad dressing, italian dressing, commercial, reduced fat
35. Salad dressing, russian dressing, low calorie
36. Salad dressing, thousand island dressing, reduced fat
37. Sandwich spread, with chopped pickle, regular, unspecified oils
38. Shortening, household, soybean (partially hydrogenated)-cottonseed (partially hydrogenated)
39. Oil, soybean, salad or cooking, (partially hydrogenated)
40. Oil, rice bran
41. Oil, wheat germ
42. Oil, peanut, salad or cooking
43. Oil, soybean, salad or cooking
44. Oil, coconut
45. Oil, olive, salad or cooking
46. Oil, palm
47. Oil, sesame, salad or cooking
48. Oil, sunflower, linoleic (less than 60%)
49. Margarine, regular, hard, soybean (hydrogenated)
50. Salad dressing, italian dressing, commercial, regular
51. Oil, cocoa butter

52. Oil, cottonseed, salad or cooking
53. Oil, sunflower, linoleic, (approx. 65%)
54. Oil, safflower, salad or cooking, linoleic, (over 70%)
55. Oil, safflower, salad or cooking, high oleic (primary safflower oil of commerce)
56. Vegetable oil, palm kernel
57. Oil, poppyseed
58. Oil, tomatoseed
59. Oil, teaseed
60. Oil, grapeseed
61. Oil, corn, industrial and retail, all purpose salad or cooking
62. Oil, walnut
63. Oil, almond
64. Oil, apricot kernel
65. Oil, hazelnut
66. Oil, babassu
67. Oil, sheanut
68. Oil, cupu assu
69. Fat, chicken
70. Oil, soybean, salad or cooking, (partially hydrogenated) and cottonseed
71. Shortening, household, lard and vegetable oil
72. Oil, sunflower, linoleic, (partially hydrogenated)
73. Shortening bread, soybean (hydrogenated) and cottonseed
74. Shortening cake mix, soybean (hydrogenated) and cottonseed (hydrogenated)
75. Shortening industrial, lard and vegetable oil
76. Shortening frying (heavy duty), beef tallow and cottonseed
77. Shortening confectionery, coconut (hydrogenated) and or palm kernel (hydrogenated)
78. Shortening industrial, soybean (hydrogenated) and cottonseed
79. Shortening frying (heavy duty), palm (hydrogenated)

80. Shortening household soybean (hydrogenated) and palm
81. Shortening frying (heavy duty), soybean (hydrogenated), linoleic (less than 1%)
82. Shortening, confectionery, fractionated palm
83. Oil, nutmeg butter
84. Oil, ucuhuba butter
85. Fat, turkey
86. Fat, goose
87. Salad dressing, mayonnaise, light
88. Oil, industrial, coconut, principal uses candy coatings, oil sprays, roasting nuts
89. Oil, industrial, soy (partially hydrogenated), principal uses popcorn and flavoring vegetables
90. Shortening, industrial, soy (partially hydrogenated), pourable liquid fry shortening
91. Oil, industrial, soy, refined, for woks and light frying
92. Oil, industrial, soy (partially hydrogenated), multiuse for non-dairy butter flavor
93. Oil, industrial, soy (partially hydrogenated), all purpose
94. Oil, industrial, soy (partially hydrogenated) and soy (winterized), pourable clear fry
95. Oil, industrial, soy (partially hydrogenated) and cottonseed, principal use as a tortilla shortening
96. Oil, industrial, palm kernel, confection fat, uses similar to high quality cocoa butter
97. Oil, industrial, palm kernel (hydrogenated), confection fat, uses similar to 95 degree hard butter
98. Oil, industrial, palm kernel (hydrogenated), confection fat, intermediate grade product
99. Oil, industrial, coconut, confection fat, typical basis for ice cream coatings
100. Oil, industrial, palm kernel (hydrogenated) , used for whipped toppings, non-dairy

101. Oil, industrial, coconut (hydrogenated), used for whipped toppings and coffee whiteners
102. Oil, industrial, palm and palm kernel, filling fat (non-hydrogenated)
103. Oil, industrial, palm kernel (hydrogenated), filling fat
104. Oil, industrial, soy (partially hydrogenated), palm, principal uses icings and fillings
105. Shortening, industrial, soy (partially hydrogenated) for baking and confections
106. Oil, vegetable, soybean, refined
107. Soup, cream of celery, canned, condensed
108. Soup, cream of mushroom, canned, condensed
109. Sauce, ready-to-serve, pepper, TABASCO
110. CAMPBELL'S, Cream of Mushroom Soup, condensed
111. Soup, cream of asparagus, canned, prepared with equal volume water
112. Soup, cream of celery, canned, prepared with equal volume water
113. Soup, chicken gumbo, canned, prepared with equal volume water
114. Soup, cream of potato, canned, prepared with equal volume water
115. Soup, turkey vegetable, canned, prepared with equal volume water
116. Soup, tomato bisque, canned, prepared with equal volume water
117. Gravy, HEINZ Home Style, savory beef
118. Cereals, corn grits, yellow, regular, quick, enriched, cooked with water, with salt
119. Cereals, CREAM OF RICE, cooked with water, with salt
120. Apples, raw, with skin (Includes foods for USDA's Food Distribution Program)
121. Apples, raw, without skin
122. Apples, raw, without skin, cooked, microwave
123. Apples, canned, sweetened, sliced, drained, heated
124. Apples, dehydrated (low moisture), sulfured, uncooked
125. Apples, dehydrated (low moisture), sulfured, stewed
126. Apples, dried, sulfured, stewed, without added sugar

127. Apples, frozen, unsweetened, unheated (Includes foods for USDA's Food Distribution Program)
128. Apples, frozen, unsweetened, heated (Includes foods for USDA's Food Distribution Program)
129. Applesauce, canned, unsweetened, without added ascorbic acid (Includes foods for USDA's Food Distribution Program)
130. Applesauce, canned, sweetened, without salt
131. Apricots, raw
132. Apricots, canned, water pack, with skin, solids and liquids
133. Apricots, canned, water pack, without skin, solids and liquids
134. Apricots, canned, juice pack, with skin, solids and liquids
135. Apricots, canned, extra light syrup pack, with skin, solids and liquids (Includes foods for USDA's Food Distribution Program)
136. Apricots, canned, light syrup pack, with skin, solids and liquids
137. Apricots, canned, heavy syrup pack, with skin, solids and liquids
138. Apricots, dried, sulfured, stewed, without added sugar
139. Bananas, raw
140. Blueberries, raw
141. Blueberries, canned, heavy syrup, solids and liquids
142. Blueberries, frozen, unsweetened (Includes foods for USDA's Food Distribution Program)
143. Blueberries, frozen, sweetened
144. Breadfruit, raw
145. Cherries, sweet, raw
146. Cranberries, raw
147. Elderberries, raw
148. Figs, raw
149. Figs, canned, water pack, solids and liquids
150. Figs, canned, light syrup pack, solids and liquids

151. Figs, canned, heavy syrup pack, solids and liquids
152. Figs, dried, stewed
153. Grapefruit, raw, pink and red and white, all areas
154. Grapefruit, raw, pink and red, all areas
155. Grapefruit, raw, pink and red, California and Arizona
156. Grapefruit, raw, pink and red, Florida
157. Grapefruit, raw, white, all areas
158. Grapefruit, raw, white, Florida
159. Grapefruit, sections, canned, water pack, solids and liquids
160. Grapefruit, sections, canned, juice pack, solids and liquids
161. Grapefruit, sections, canned, light syrup pack, solids and liquids
162. Grapes, american type (slip skin), raw
163. Grapes, red or green (European type, such as Thompson seedless), raw
164. Grapes, canned, thompson seedless, water pack, solids and liquids
165. Grapes, canned, thompson seedless, heavy syrup pack, solids and liquids
166. Grape juice, canned or bottled, unsweetened, without added ascorbic acid
167. Guavas, strawberry, raw
168. Guava sauce, cooked
169. Kiwifruit, green, raw
170. Longans, raw
171. Loquats, raw
172. Mangos, raw
173. Melons, cantaloupe, raw
174. Melons, honeydew, raw
175. Nectarines, raw
176. Olives, ripe, canned (small-extra large)
177. Olives, ripe, canned (jumbo-super colossal)
178. Oranges, raw, all commercial varieties
179. Oranges, raw, California, valencias

180. Oranges, raw, navels (Includes foods for USDA's Food Distribution Program)
181. Oranges, raw, Florida
182. Oranges, raw, with peel
183. Orange juice, raw (Includes foods for USDA's Food Distribution Program)
184. Orange juice, canned, unsweetened
185. Orange juice, chilled, includes from concentrate
186. Orange juice, chilled, includes from concentrate, with added calcium and vitamin D
187. Orange juice, chilled, includes from concentrate, with added calcium
188. Tangerines, (mandarin oranges), raw
189. Tangerines, (mandarin oranges), canned, juice pack
190. Tangerines, (mandarin oranges), canned, light syrup pack
191. Tangerine juice, raw
192. Papayas, raw
193. Peaches, yellow, raw
194. Peaches, canned, water pack, solids and liquids
195. Peaches, canned, juice pack, solids and liquids
196. Peaches, canned, extra light syrup, solids and liquids (Includes foods for USDA's Food Distribution Program)
197. Peaches, canned, light syrup pack, solids and liquids
198. Peaches, canned, heavy syrup pack, solids and liquids
199. Peaches, spiced, canned, heavy syrup pack, solids and liquids
200. Peaches, dried, sulfured, stewed, without added sugar
201. Peaches, frozen, sliced, sweetened
202. Pears, raw
203. Pears, canned, water pack, solids and liquids
204. Pears, canned, juice pack, solids and liquids
205. Pears, canned, extra light syrup pack, solids and liquids (Includes foods for USDA's Food Distribution Program)

206. Pears, canned, light syrup pack, solids and liquids
207. Pears, canned, heavy syrup pack, solids and liquids
208. Pears, dried, sulfured, stewed, without added sugar
209. Persimmons, japanese, raw
210. Pineapple, raw, all varieties
211. Pineapple, canned, water pack, solids and liquids
212. Pineapple, canned, juice pack, solids and liquids
213. Pineapple, canned, light syrup pack, solids and liquids
214. Pineapple, canned, heavy syrup pack, solids and liquids
215. Pineapple, frozen, chunks, sweetened
216. Plums, raw
217. Plums, canned, purple, water pack, solids and liquids
218. Plums, canned, purple, juice pack, solids and liquids
219. Plums, canned, purple, light syrup pack, solids and liquids
220. Plums, canned, purple, heavy syrup pack, solids and liquids
221. Sapodilla, raw
222. Sapote, mamey, raw
223. Strawberries, canned, heavy syrup pack, solids and liquids
224. Strawberries, frozen, unsweetened (Includes foods for USDA's Food Distribution Program)
225. Watermelon, raw
226. Feijoa, raw
227. Pears, asian, raw
228. Peaches, canned, heavy syrup, drained
229. Applesauce, canned, unsweetened, with added ascorbic acid
230. Applesauce, canned, sweetened, with salt
231. Pears, raw, bartlett (Includes foods for USDA's Food Distribution Program)
232. Pears, raw, red anjou
233. Pears, raw, bosc (Includes foods for USDA's Food Distribution Program)

234. Pears, raw, green anjou (Includes foods for USDA's Food Distribution Program)
235. Apples, raw, red delicious, with skin (Includes foods for USDA's Food Distribution Program)
236. Apples, raw, golden delicious, with skin
237. Orange juice, chilled, includes from concentrate, with added calcium and vitamins A, D, E
238. Grape juice, canned or bottled, unsweetened, with added ascorbic acid and calcium
239. Beans, snap, green, canned, regular pack, solids and liquids
240. Beans, snap, green, canned, regular pack, drained solids
241. Beans, snap, canned, all styles, seasoned, solids and liquids
242. Beans, snap, green, frozen, cooked, boiled, drained without salt
243. Beets, canned, regular pack, solids and liquids
244. Beets, canned, drained solids
245. Cabbage, cooked, boiled, drained, without salt
246. Cabbage, red, raw
247. Cabbage, chinese (pe-tsai), raw
248. Carrots, raw
249. Carrots, canned, regular pack, solids and liquids
250. Carrots, canned, regular pack, drained solids
251. Carrots, frozen, unprepared (Includes foods for USDA's Food Distribution Program)
252. Carrots, frozen, cooked, boiled, drained, without salt
253. Celery, raw
254. Celery, cooked, boiled, drained, without salt
255. Celtuce, raw
256. Chayote, fruit, raw
257. Chayote, fruit, cooked, boiled, drained, without salt
258. Cucumber, with peel, raw

259. Cucumber, peeled, raw
260. Eggplant, cooked, boiled, drained, without salt
261. Endive, raw
262. Gourd, white-flowered (calabash), raw
263. Gourd, white-flowered (calabash), cooked, boiled, drained, without salt
264. Leeks, (bulb and lower leaf-portion), cooked, boiled, drained, without salt
265. Lettuce, iceberg (includes crisphead types), raw
266. Onions, raw
267. Onions, cooked, boiled, drained, without salt
268. Onions, canned, solids and liquids
269. Onions, frozen, chopped, unprepared
270. Onions, frozen, chopped, cooked, boiled, drained, without salt
271. Onions, frozen, whole, unprepared
272. Onions, frozen, whole, cooked, boiled, drained, without salt
273. Onions, sweet, raw
274. Peppers, hot chili, green, canned, pods, excluding seeds, solids and liquids
275. Peppers, sweet, green, cooked, boiled, drained, without salt
276. Peppers, sweet, green, canned, solids and liquids
277. Peppers, sweet, green, frozen, chopped, unprepared
278. Peppers, sweet, green, frozen, chopped, boiled, drained, without salt
279. Peppers, sweet, green, sauteed
280. Pumpkin, raw
281. Pumpkin, cooked, boiled, drained, without salt
282. Radishes, raw
283. Radishes, oriental, raw
284. Radishes, oriental, cooked, boiled, drained, without salt
285. Sesbania flower, raw
286. Squash, summer, crookneck and straightneck, raw

287. Squash, summer, crookneck and straightneck, cooked, boiled, drained, without salt
288. Squash, summer, crookneck and straightneck, canned, drained, solid, without salt
289. Squash, summer, crookneck and straightneck, frozen, unprepared
290. Squash, summer, crookneck and straightneck, frozen, cooked, boiled, drained, without salt
291. Squash, summer, scallop, raw
292. Squash, summer, scallop, cooked, boiled, drained, without salt
293. Squash, summer, zucchini, includes skin, raw
294. Squash, summer, zucchini, includes skin, cooked, boiled, drained, without salt
295. Squash, summer, zucchini, includes skin, frozen, unprepared
296. Squash, summer, zucchini, includes skin, frozen, cooked, boiled, drained, without salt
297. Squash, summer, zucchini, italian style, canned
298. Squash, winter, acorn, raw
299. Squash, winter, acorn, cooked, baked, without salt
300. Squash, winter, butternut, raw
301. Squash, winter, butternut, cooked, baked, without salt
302. Squash, winter, spaghetti, raw
303. Squash, winter, spaghetti, cooked, boiled, drained, or baked, without salt
304. Tomatoes, green, raw
305. Tomatoes, red, ripe, raw, year round average
306. Tomatoes, red, ripe, cooked
307. Tomatoes, red, ripe, canned, packed in tomato juice
308. Tomatoes, red, ripe, canned, stewed
309. Tomato products, canned, sauce
310. Tomato products, canned, sauce, with mushrooms
311. Tomato products, canned, sauce, with tomato tidbits

312. Turnips, raw
313. Turnips, frozen, unprepared
314. Turnip greens, cooked, boiled, drained, without salt
315. Vegetable juice cocktail, canned
316. Vegetables, mixed, canned, solids and liquids
317. Vegetable juice cocktail, low sodium, canned
318. Yambean (jicama), cooked, boiled, drained, without salt
319. Beets, harvard, canned, solids and liquids
320. Beets, pickled, canned, solids and liquids
321. Peppers, jalapeno, canned, solids and liquids
322. Radishes, white icicle, raw
323. Squash, summer, all varieties, raw
324. Squash, summer, all varieties, cooked, boiled, drained, without salt
325. Sweet potato, canned, syrup pack, solids and liquids
326. Tomato products, canned, sauce, spanish style
327. Tomatoes, orange, raw
328. Tomatoes, yellow, raw
329. Beans, snap, green, canned, no salt added, solids and liquids
330. Beans, snap, yellow, canned, regular pack, solids and liquids
331. Beans, snap, yellow, canned, no salt added, solids and liquids
332. Beans, snap, green, canned, no salt added, drained solids
333. Beans, snap, yellow, frozen, cooked, boiled, drained, with salt
334. Beets, canned, no salt added, solids and liquids
335. Cabbage, common, cooked, boiled, drained, with salt
336. Carrots, canned, no salt added, solids and liquids
337. Carrots, canned, no salt added, drained solids
338. Carrots, frozen, cooked, boiled, drained, with salt
339. Celery, cooked, boiled, drained, with salt
340. Gourd, white-flowered (calabash), cooked, boiled, drained, with salt

341. Leeks, (bulb and lower leaf-portion), cooked, boiled, drained, with salt
342. Onions, cooked, boiled, drained, with salt
343. Onions, frozen, chopped, cooked, boiled, drained, with salt
344. Onions, frozen, whole, cooked, boiled, drained, with salt
345. Peppers, sweet, red, raw
346. Peppers, sweet, green, cooked, boiled, drained, with salt
347. Peppers, sweet, red, cooked, boiled, drained, without salt
348. Peppers, sweet, red, cooked, boiled, drained, with salt
349. Peppers, sweet, green, frozen, chopped, cooked, boiled, drained, with salt
350. Pumpkin, cooked, boiled, drained, with salt
351. Radishes, oriental, cooked, boiled, drained, with salt
352. Squash, summer, all varieties, cooked, boiled, drained, with salt
353. Squash, summer, crookneck and straightneck, cooked, boiled, drained, with salt
354. Squash, summer, crookneck and straightneck, frozen, cooked, boiled, drained, with salt
355. Squash, summer, scallop, cooked, boiled, drained, with salt
356. Squash, summer, zucchini, includes skin, cooked, boiled, drained, with salt
357. Squash, summer, zucchini, includes skin, frozen, cooked, boiled, drained, with salt
358. Squash, winter, acorn, cooked, baked, with salt
359. Squash, winter, spaghetti, cooked, boiled, drained, or baked, with salt
360. Tomatoes, red, ripe, cooked, with salt
361. Tomatoes, red, ripe, canned, packed in tomato juice, no salt added
362. Tomato juice, canned, without salt added
363. Yambean (jicama), cooked, boiled, drained, with salt
364. Peppers, sweet, red, canned, solids and liquids
365. Peppers, sweet, red, frozen, chopped, unprepared
366. Peppers, sweet, red, frozen, chopped, boiled, drained, without salt
367. Peppers, sweet, red, frozen, chopped, boiled, drained, with salt

368. Peppers, sweet, red, sauteed
369. Sesbania flower, cooked, steamed, with salt
370. Beans, snap, yellow, canned, regular pack, drained solids
371. Beans, snap, yellow, canned, no salt added, drained solids
372. Catsup
373. Pickles, cucumber, dill or kosher dill
374. Pickles, cucumber, sweet (includes bread and butter pickles)
375. Pimento, canned
376. Pickle relish, sweet
377. Pickles, cucumber, sour, low sodium
378. Pickles, cucumber, dill, reduced sodium
379. Pickles, cucumber, sweet, low sodium (includes bread and butter pickles)
380. Catsup, low sodium
381. Peppers, sweet, yellow, raw
382. Radicchio, raw
383. Nopales, raw
384. Peppers, chili, green, canned
385. Peppers, hungarian, raw
386. Nuts, coconut water (liquid from coconuts)
387. Nuts, chestnuts, japanese, boiled and steamed
388. Alcoholic beverage, beer, regular, all
389. Alcoholic beverage, creme de menthe, 72 proof
390. Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 80 proof
391. Alcoholic beverage, distilled, rum, 80 proof
392. Alcoholic beverage, distilled, vodka, 80 proof
393. Beverages, almond milk, chocolate, ready-to-drink
394. Beverages, carbonated, club soda
395. Carbonated beverage, cream soda
396. Beverages, carbonated, grape soda

397. Beverages, carbonated, orange
398. Beverages, carbonated, pepper-type, contains caffeine
399. Beverages, carbonated, tonic water
400. Beverages, Clam and tomato juice, canned
401. Beverages, coffee, brewed, prepared with tap water, decaffeinated
402. Beverages, coffee, brewed, prepared with tap water
403. Cranberry juice cocktail, bottled, low calorie, with calcium, saccharin and corn sweetener
404. Beverages, tea, black, brewed, prepared with tap water, decaffeinated
405. Beverages, tea, black, brewed, prepared with tap water
406. Beverages, water, bottled, PERRIER
407. Beverages, water, bottled, POLAND SPRING
408. Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 94 proof
409. Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 100 proof
410. Beverages, tea, black, brewed, prepared with distilled water
411. Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 86 proof
412. Alcoholic beverage, distilled, all (gin, rum, vodka, whiskey) 90 proof
413. Mollusks, clam, mixed species, canned, liquid
414. Gelatin desserts, dry mix, prepared with water
415. Puddings, lemon, dry mix, regular
416. Puddings, banana, dry mix, regular, with added oil
417. Puddings, vanilla, dry mix, regular, with added oil
418. Hominy, canned, yellow
419. KFC, Coleslaw
420. Soup, egg drop, Chinese restaurant
421. APPLEBEE'S, coleslaw
422. CRACKER BARREL, coleslaw
423. Tomato sauce, canned, no salt added
424. Babyfood, grape juice, no sugar, canned

B. LIST OF FOODS ESTIMATED WITH BAD ACCURACY (ERROR > 50MG PHE PER 100G OF FOOD) (K=4)

1. Butter, salted
2. Cream substitute, liquid, with hydrogenated vegetable oil and soy protein
3. Cream substitute, liquid, with lauric acid oil and sodium caseinate
4. Dessert topping, pressurized
5. Cream substitute, flavored, powdered
6. Salt, table
7. Babyfood, juice treats, fruit medley, toddler
8. Babyfood, snack, GERBER GRADUATE FRUIT STRIPS, Real Fruit Bars
9. Babyfood, vegetables, green beans, strained
10. Babyfood, cereal, oatmeal, with applesauce and bananas, strained
11. Babyfood, cereal, oatmeal, with applesauce and bananas, junior, fortified
12. Babyfood, vegetables, mix vegetables junior
13. Babyfood, mashed cheddar potatoes and broccoli, toddlers
14. Salad dressing, thousand island, commercial, regular
15. Salad dressing, mayonnaise, regular
16. Salad dressing, mayonnaise, soybean and safflower oil, with salt
17. Salad dressing, mayonnaise, imitation, soybean
18. Salad dressing, mayonnaise, imitation, soybean without cholesterol
19. Salad dressing, french, home recipe
20. Salad dressing, home recipe, vinegar and oil
21. Sauce, ready-to-serve, pepper or hot
22. Soup, cream of chicken, canned, prepared with equal volume water

23. Cereals, CREAM OF WHEAT, regular (10 minute), cooked with water, without salt
24. Cereals, corn grits, yellow, regular and quick, enriched, cooked with water, without salt
25. Cereals, CREAM OF WHEAT, regular (10 minute), cooked with water, with salt
26. Cereals, CREAM OF WHEAT, 2 1/2 minute cook time, cooked with water, stove-top, without salt
27. Apples, dried, sulfured, uncooked
28. Apples, dried, sulfured, stewed, with added sugar
29. Apricots, canned, heavy syrup pack, without skin, solids and liquids
30. Apricots, canned, extra heavy syrup pack, without skin, solids and liquids
31. Apricots, dried, sulfured, stewed, with added sugar
32. Apricots, frozen, sweetened
33. Carambola, (starfruit), raw
34. Crabapples, raw
35. Figs, canned, extra heavy syrup pack, solids and liquids
36. Grapefruit, raw, white, California
37. Lime juice, raw
38. Peaches, canned, extra heavy syrup pack, solids and liquids
39. Peaches, dried, sulfured, stewed, with added sugar
40. Pears, canned, extra heavy syrup pack, solids and liquids
41. Pears, dried, sulfured, stewed, with added sugar
42. Persimmons, japanese, dried
43. Persimmons, native, raw
44. Pineapple, canned, extra heavy syrup pack, solids and liquids
45. Plums, canned, purple, extra heavy syrup pack, solids and liquids
46. Strawberries, raw
47. Strawberries, frozen, sweetened, sliced

48. Plantains, green, boiled
49. Cabbage, raw
50. Carrots, cooked, boiled, drained, without salt
51. Cassava, raw
52. Chicory, witloof, raw
53. Eggplant, raw
54. Escarole, cooked, boiled, drained, no salt added
55. Lettuce, butterhead (includes boston and bibb types), raw
56. Lettuce, cos or romaine, raw
57. Lettuce, green leaf, raw
58. Lettuce, red leaf, raw
59. Mountain yam, hawaii, raw
60. Onions, yellow, sauteed
61. Peppers, sweet, green, raw
62. Potatoes, canned, solids and liquids
63. Potatoes, canned, drained solids
64. Pumpkin, canned, without salt
65. Pumpkin pie mix, canned
66. Purslane, cooked, boiled, drained, without salt
67. Sauerkraut, canned, solids and liquids
68. Sesbania flower, cooked, steamed, without salt
69. Squash, winter, acorn, cooked, boiled, mashed, without salt
70. Squash, winter, butternut, frozen, cooked, boiled, without salt
71. Squash, winter, hubbard, cooked, boiled, mashed, without salt
72. Sweet potato, cooked, boiled, without skin
73. Taro, cooked, without salt
74. Tomato products, canned, sauce, with onions, green peppers, and celery
75. Turnips, cooked, boiled, drained, without salt
76. Turnip greens, canned, solids and liquids

77. Yam, cooked, boiled, drained, or baked, without salt
78. Yambean (jicama), raw
79. Beans, mung, mature seeds, sprouted, canned, drained solids
80. Squash, winter, all varieties, raw
81. Squash, winter, all varieties, cooked, baked, without salt
82. Sweet potato, canned, syrup pack, drained solids
83. Sweet potato, cooked, candied, home-prepared
84. Beans, snap, green, frozen, cooked, boiled, drained, with salt
85. Beans, snap, yellow, frozen, cooked, boiled, drained, without salt
86. Cabbage, common (danish, domestic, and pointed types), freshly harvest, raw
87. Cabbage, common (danish, domestic, and pointed types), stored, raw
88. Carrots, cooked, boiled, drained, with salt
89. Chayote, fruit, cooked, boiled, drained, with salt
90. Eggplant, cooked, boiled, drained, with salt
91. Peppers, hot chili, red, canned, excluding seeds, solids and liquids
92. Pumpkin, canned, with salt
93. Purslane, cooked, boiled, drained, with salt
94. Squash, winter, all varieties, cooked, baked, with salt
95. Squash, winter, acorn, cooked, boiled, mashed, with salt
96. Squash, winter, butternut, cooked, baked, with salt
97. Squash, winter, butternut, frozen, cooked, boiled, with salt
98. Squash, winter, hubbard, cooked, boiled, mashed, with salt
99. Sweet potato, cooked, boiled, without skin, with salt
100. Taro, cooked, with salt
101. Turnips, cooked, boiled, drained, with salt
102. Turnip greens, cooked, boiled, drained, with salt
103. Yam, cooked, boiled, drained, or baked, with salt
104. Pickles, cucumber, sour
105. Pickle relish, hamburger

106. Carrots, baby, raw
107. Nopales, cooked, without salt
108. Nuts, coconut cream, canned, sweetened
109. Alcoholic beverage, beer, light
110. Beverages, Orange drink, breakfast type, with juice and pulp, frozen concentrate
111. Beverages, tea, instant, lemon, with added ascorbic acid
112. Noodles, chinese, cellophane or long rice (mung beans), dehydrated
113. Fruit syrup
114. Puddings, tapioca, dry mix
115. Puddings, vanilla, dry mix, regular
116. Frostings, chocolate, creamy, ready-to-eat
117. Frostings, cream cheese-flavor, ready-to-eat
118. Frostings, chocolate, creamy, dry mix
119. Frostings, chocolate, creamy, dry mix, prepared with butter
120. Frostings, vanilla, creamy, dry mix
121. Honey
122. Jams and preserves
123. Marmalade, orange
124. Pie fillings, canned, cherry
125. Puddings, banana, dry mix, regular
126. Puddings, lemon, dry mix, instant
127. Toppings, butterscotch or caramel
128. Frostings, vanilla, creamy, dry mix, prepared with margarine
129. Frostings, chocolate, creamy, dry mix, prepared with margarine
130. Puddings, lemon, dry mix, regular, with added oil, potassium, sodium
131. Puddings, tapioca, dry mix, with no added salt
132. Syrup, NESTLE, chocolate
133. Arrowroot flour
134. Cornstarch

135. Hominy, canned, white
136. Tapioca, pearl, dry
137. POPEYES, Coleslaw
138. Agave, raw (Southwest)
139. DENNY'S, coleslaw