**Analysis:**

I used TensorFlow to create a model which predicts the call values using the numerical values in the CSVs. My model had an accuracy of 99.2%, so there will be a small disparity in the accuracy and error calculations. This can be seen by comparing the original Cycle 1 accuracy and error with the code’s Cycle 1 accuracy and error. The model could produce a better accuracy if more data was presented.

I recommend that the research project should progress with the massively parallel DNA sequencing method used for sequencing\_data\_biochem1.csv because the average accuracy was 94.6% and the average error was 5.4%. The massively parallel DNA sequencing method used for sequencing\_data\_biochem2.csv had an average accuracy of 70.1% and an average error of 29.9%. This method didn’t produce any of the correct values for reference base Thymine, and produced double the N values for the call column compared to the first method. The data below displays a small description from what was produced by the code. The code’s output for each csv can be evaluated in the **Analysis Data from Code** section below.

processed\_sequencing\_data\_biochem1.csv:

The original accuracy for cycle 1 was 96.1%, with an error of 3.9%

Cycle 1:

Accuracy: 95.9%

Error: 4.1%

Amount of N values: 21

Cycle 2:

Accuracy: 93.3%

Error: 6.7%

Amount of N values: 22

Cycle 2 had a larger disparity with Adenine compared to Cycle 1.

In Cycle 1:

4 Adenine references displayed Guanine, 1 Adenine references displayed Thymine,

4 Adenine references displayed Cytosine.

In Cycle 2:

16 Adenine references displayed Guanine, 12 Adenine references displayed Thymine,

9 Adenine references displayed Cytosine.

processed\_sequencing\_data\_biochem2\_revised.csv:

All Thymine values were incorrect.

None of the other bases were incorrectly displaying Thymine.

Cycle 1:

Accuracy: 71%

Error: 29%

Amount of N values: 43

Cycle 2:

Accuracy: 69.2%

Error: 30.8%

Amount of N values: 43

**Analysis Data from Code:**

processed\_sequencing\_data\_biochem2\_revised.csv

Cycle 1:

ref\_1 call\_1 count

0 A A 211

1 A C 4

2 A G 4

3 A N 3

4 A T 1

5 C A 1

6 C C 253

7 C G 1

8 C N 6

9 G C 1

10 G G 244

11 G N 6

12 G T 1

13 T C 2

14 T G 5

15 T N 6

16 T T 251

Accuracy: 95.89999999999999 %

Error: 4.100000000000003 %

Amount of N values: 21

Cycle 2:

ref\_2 call\_2 count

0 A A 243

1 A C 9

2 A G 16

3 A N 6

4 A T 12

5 C C 226

6 C G 1

7 C N 5

8 C T 1

9 G C 2

10 G G 246

11 G N 8

12 G T 1

13 T C 2

14 T G 1

15 T N 3

16 T T 218

Accuracy: 93.30000000000001 %

Error: 6.699999999999995 %

Amount of N values: 22

Average Accuracy of Method: 94.6 %

Average Error of Method: 5.400000000000006 %

processed\_sequencing\_data\_biochem2\_revised.csv

Cycle 1:

ref\_1 call\_1 count

0 A A 241

1 A C 4

2 A G 8

3 A N 7

4 C A 2

5 C C 207

6 C G 9

7 C N 4

8 G A 2

9 G C 3

10 G G 262

11 G N 7

12 T A 54

13 T C 78

14 T G 87

15 T N 25

Accuracy: 71.0 %

Error: 29.000000000000004 %

Amount of N values: 43

Cycle 2:

ref\_2 call\_2 count

0 A A 225

1 A C 3

2 A G 6

3 A N 8

4 C A 5

5 C C 234

6 C G 8

7 C N 6

8 G A 1

9 G G 233

10 G N 4

11 T A 70

12 T C 80

13 T G 92

14 T N 25

Accuracy: 69.19999999999999 %

Error: 30.800000000000004 %

Amount of N values: 43

Average Accuracy of Method: 70.1 %

Average Error of Method: 29.900000000000006 %