

COMP9444 Assignment 1

Charles Wyatt z5194905

Contents

Part 1.....	2
Question 1.....	2
Confusion Matrix and Final Accuracy of NetLin.....	2
Question 2.....	2
Full network –	2
Confusion Matrix and Accuracy of full network with 250 hidden nodes	2
Question 3.....	3
Confusion Matrix and Accuracy of convolution network	3
Question 4.....	3
Part 2.....	4
Question 2.....	4
Question 4.....	5
Question 5.....	5
Question 7.....	6
Part 3.....	7
Part 4.....	7
Question 1.....	7
Question 2.....	8
Question 3.....	8
Question 4.....	8
Question 5.....	9
Question 6.....	9
Appendices.....	10
Appendix A.....	10
Appendix B.....	28

Part 1

Question 1

Confusion Matrix and Final Accuracy of NetLin

```
[[766.  6.  9. 13. 29. 63.  2. 62. 32. 18.]
 [  6. 668. 108. 19. 28. 23. 59. 12. 26. 51.]
 [  8. 63. 687. 25. 26. 21. 48. 37. 45. 40.]
 [  4. 37. 56. 761. 15. 54. 15. 18. 29. 11.]
 [ 61. 52. 80. 19. 623. 20. 32. 36. 22. 55.]
 [  7. 27. 124. 17. 19. 725. 27. 10. 34. 10.]
 [  5. 24. 146. 10. 23. 23. 724. 21.  9. 15.]
 [ 18. 28. 28. 11. 82. 18. 56. 619. 92. 48.]
 [ 11. 38. 92. 40.  8. 31. 44.  6. 707. 23.]
 [  9. 50. 91.  3. 54. 29. 20. 32. 39. 673.]]
```

Test set: Average loss: 1.0091, Accuracy: 6953/10000 (70%)

Question 2

Full network –

Number of Hidden nodes	Accuracy
20	73%
30	76%
40	79%
50	80%
100	83%
150	84%
200	84%
250	84.97%
300	85.15%
350	85.4%
400	85.51%

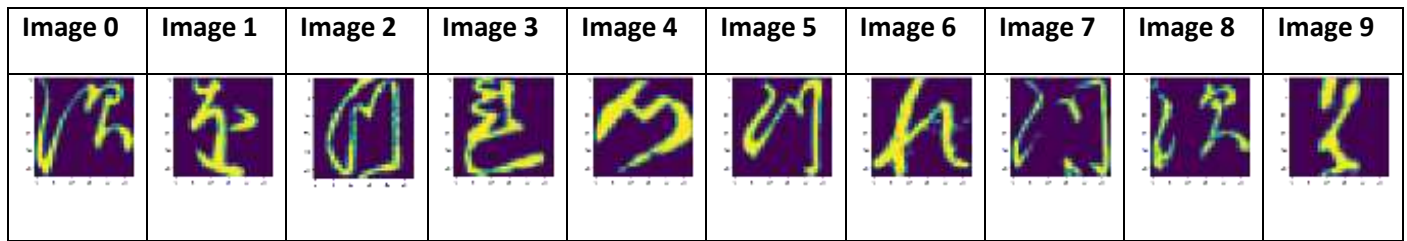
250 is chosen as the best value since greater values showed no material improvement and took longer to train.

Confusion Matrix and Accuracy of full network with 250 hidden nodes

```
[[860.  4.  1.  5. 26. 25.  3. 41. 29.  6.]
 [  5. 816. 30.  6. 19.  9. 57.  8. 23. 27.]
 [  8. 16. 837. 36. 14. 17. 27. 10. 18. 17.]
 [  3.  7. 32. 918.  2. 14.  5.  5.  7.  7.]
 [ 43. 25. 16.  5. 818.  6. 31. 20. 18. 18.]
 [ 11. 10. 68.  9. 12. 839. 28.  2. 14.  7.]
 [  3. 15. 39.  8. 12.  5. 903.  8.  2.  5.]
 [ 16. 15. 16.  4. 28. 11. 27. 824. 24. 35.]
 [ 14. 28. 24. 47.  3.  6. 31.  4. 837.  6.]
 [  3. 15. 40.  6. 35.  4. 21. 19. 12. 845.]]
```

Test set: Average loss: 0.4951, Accuracy: 8497/10000 (85%)

Question 3



Confusion Matrix and Accuracy of convolution network

```
[[960.  3.  1.  1. 22.  2.  0.  4.  4.  3.]
 [  2. 938.  3.  0.  2.  1. 32.  5.  5. 12.]
 [10.  5. 915. 18.  5. 13. 14. 11.  4.  5.]
 [  1.  0. 15. 962.  1.  8.  5.  3.  3.  2.]
 [23. 10.  2.  5. 937.  4.  6.  6.  4.  3.]
 [  4.  9. 44.  5.  1. 916. 12.  1.  3.  5.]
 [  4.  2. 15.  1.  4.  3. 962.  4.  0.  5.]
 [  1.  7.  4.  0.  2.  0.  4. 961.  5. 16.]
 [  6.  9.  7.  6.  9.  3.  5.  2. 951.  2.]
 [  7.  8.  6.  0.  3.  0. 10.  4.  5. 957.]]
```

Test set: Average loss: 0.2225, Accuracy: 9459/10000 (95%)

Question 4

- A) Clearly, the convolution network is the most accurate model at a 95% accuracy. Since the convolution network has more independent parameters it is a much more flexible, and so can fit the complexities of the image. The full network is the second most accurate network at 85% and its predictive power becomes stronger by increasing the number of hidden nodes. However, the accuracy of the full network plateaus at 250. Lastly, the worst model was the linear network with an accuracy of only 70%
- B) In this case, the number of independent parameters scales with the predictive capabilities of the model.

Note, all independent parameters were calculated through the torchsummary function.

Model	Independent Parameters
Linear Network	7 850
Full Network	198 760
Convolution Network	313 108

- C) See the following table for each model and use the examples of the characters as a reference

Model	Confusion Matrix	Most likely mistaken	Reasons
Linear Network	<pre>[[766. 6. 9. 13. 29. 63. 2. 62. 32. 18.] [6. 668. 108. 19. 28. 23. 59. 12. 26. 51.] [8. 63. 687. 25. 26. 21. 48. 37. 45. 40.] [4. 37. 56. 761. 15. 54. 15. 18. 29. 11.] [61. 52. 80. 19. 623. 20. 32. 36. 22. 55.] [7. 27. 124. 17. 19. 725. 27. 10. 34. 10.] [5. 24. 146. 10. 23. 23. 724. 21. 9. 15.] [18. 28. 28. 11. 82. 18. 56. 619. 92. 48.] [11. 38. 92. 40. 8. 31. 44. 6. 707. 23.] [9. 50. 91. 3. 54. 29. 20. 32. 39. 673.]]</pre>	<ul style="list-style-type: none"> Images 2 and 1. Images 4 and 2. Images 5 and 2 Images 2 and 8 Images 2 and 9 	<ul style="list-style-type: none"> Image 2 and image 1 were the most mistaken, perhaps due to the similarities in the left-hand side of the image. Both are quite thin characters with lots of "curls" and so appear similar to a

			simple linear network
Full Network	<pre>[[860. 4. 1. 5. 26. 25. 3. 41. 29. 6.] [5. 816. 30. 6. 19. 9. 57. 8. 23. 27.] [8. 16. 837. 36. 14. 17. 27. 10. 18. 17.] [3. 7. 32. 918. 2. 14. 5. 5. 7. 7.] [43. 25. 16. 5. 818. 6. 31. 20. 18. 18.] [11. 10. 68. 9. 12. 839. 28. 2. 14. 7.] [3. 15. 39. 8. 12. 5. 903. 8. 2. 5.] [16. 15. 16. 4. 28. 11. 27. 824. 24. 35.] [14. 28. 24. 47. 3. 6. 31. 4. 837. 6.] [3. 15. 40. 6. 35. 4. 21. 19. 12. 845.]]</pre>	<ul style="list-style-type: none"> Image 2 and image 5 Image 2 and image 9 Image 7 and image 0 Image 1 and image 6 	Image 2 and 5 were the most mistaken. The right side of the characters appear very similar, with a long, thin vertical line.
Convolution Network	<pre>[[960. 3. 1. 1. 22. 2. 0. 4. 4. 3.] [2. 938. 3. 0. 2. 1. 32. 5. 5. 12.] [10. 5. 915. 18. 5. 13. 14. 11. 4. 5.] [1. 0. 15. 962. 1. 8. 5. 3. 3. 2.] [23. 10. 2. 5. 937. 4. 6. 6. 4. 3.] [4. 9. 44. 5. 1. 916. 12. 1. 3. 5.] [4. 2. 15. 1. 4. 3. 962. 4. 0. 5.] [1. 7. 4. 0. 2. 0. 4. 961. 5. 16.] [6. 9. 7. 6. 9. 3. 5. 2. 951. 2.] [7. 8. 6. 0. 3. 0. 10. 4. 5. 957.]]</pre> <p>Test set: Average loss: 0.2225, Accuracy: 9459/10000 (95%)</p>	<ul style="list-style-type: none"> Image 2 and image 5 	There were not many mistakes here. Again, Image 5 and 2 were the most mistaken. Perhaps for the same reasons as above

Part 2

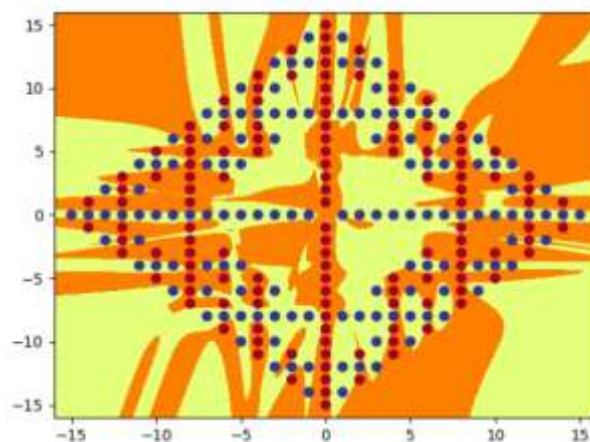
Question 2

- The initial weights used were $\frac{1}{\sqrt{\text{hidden_nodes}}}$
- The learning rate used as 0.001.

The first model that was successfully trained using these parameters was the model with 20 hidden nodes. The results from the tested models can be shown below.

Hidden nodes	Accuracy	Epochs Needed
1	~ 56%	23 100
5	~69%	21 200
10	~ 87%	41 900
15	~ 91%	110 100
20	100%	90 000

Plot of output two layer 20 hidden nodes.



Total independent Parameters:

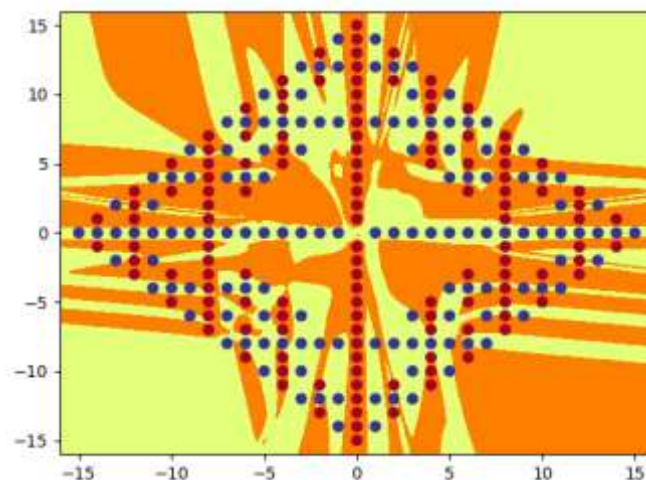
Question 4

- The initial weights used were $\frac{1}{\sqrt{\text{hidden_nodes}}}$
- The learning rate used as 0.001.

The first model that was successfully trained using these parameters was the model with 30 hidden nodes. The results from the tested models can be shown below.

Hidden nodes	Accuracy	Epochs Needed
10	~80%	200 000
20	~90%	200 000
30	100%	96 000

Output Plot



See Appendix A for the plots of all the hidden units

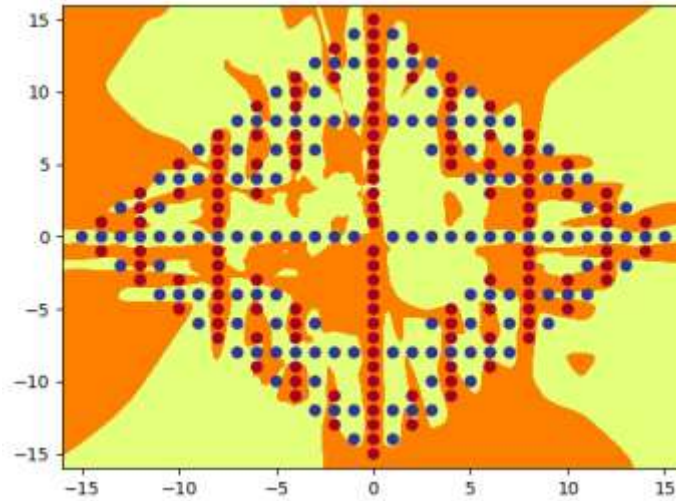
Total Independent Parameters:

1981

Question 5

Number of nodes	Accuracy	Epochs Needed
10	89%	25 000
15	96%	42 500
17	99%	53 700
20	100%	20 100

20 node output plot



See Appendix B for the plots of all the Hidden Nodes

Total Independent Parameters

563

Question 7

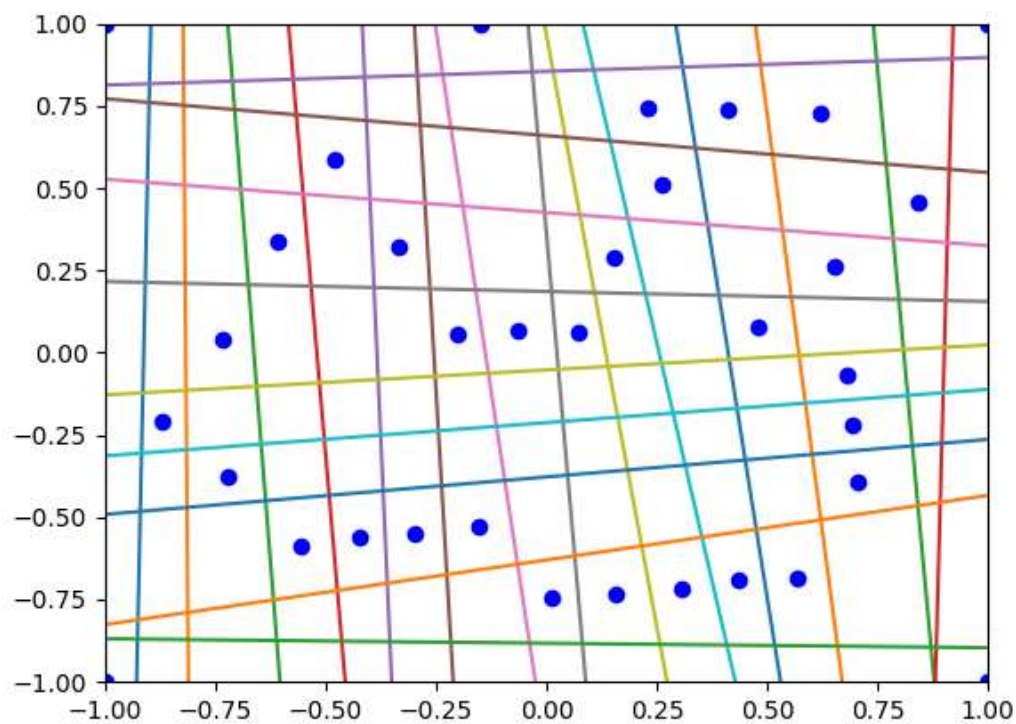
a)

Model	Hidden Nodes	Independent Parameters	Epochs
Full2Net	20	501	90 000
Full3Net	30	1981	96 000
DenseNet	20	563	20 100

Interestingly, the Full2Net required less hidden nodes than the Full3Net model, perhaps the added complexity led to some overfitting by the Full3Net model. DenseNet was clearly the best model with the least number of epochs and a similar number of independent parameters to the Full2Net.

- b) The nodes in the first layer of both Full3Net and DenseNet are linear lines. For both networks, many of the first nodes are horizontal and vertical. This is because many dots from the same class lie on these vertical and horizontal lines such as the vertical line of red dots at 0 and the horizontal line of blue dots at 0. Therefore, these nodes will be able to classify these points easily. Other lines in the first layer capture some of the diagonal elements of the fractal. In the second layer the complexity of the boundary increases. For the dense net the complexity of the boundary increases more than the 3-layer full net since it includes the skip connections. The 3-layer full network boundary again becomes more complex at the 3rd hidden layer.
- c) There were not obvious qualitative differences between the 3 output plots. However, the dense plot was able to recognize the vertical features of the red dots more effectively than the other two. The 3 layer has some diagonal areas which seem to account for the alternating blue / red pattern on the sides of the fractal, where the dense plot has instead made blue “pocket” areas as seen in the bottom right-hand corner.

Part 3



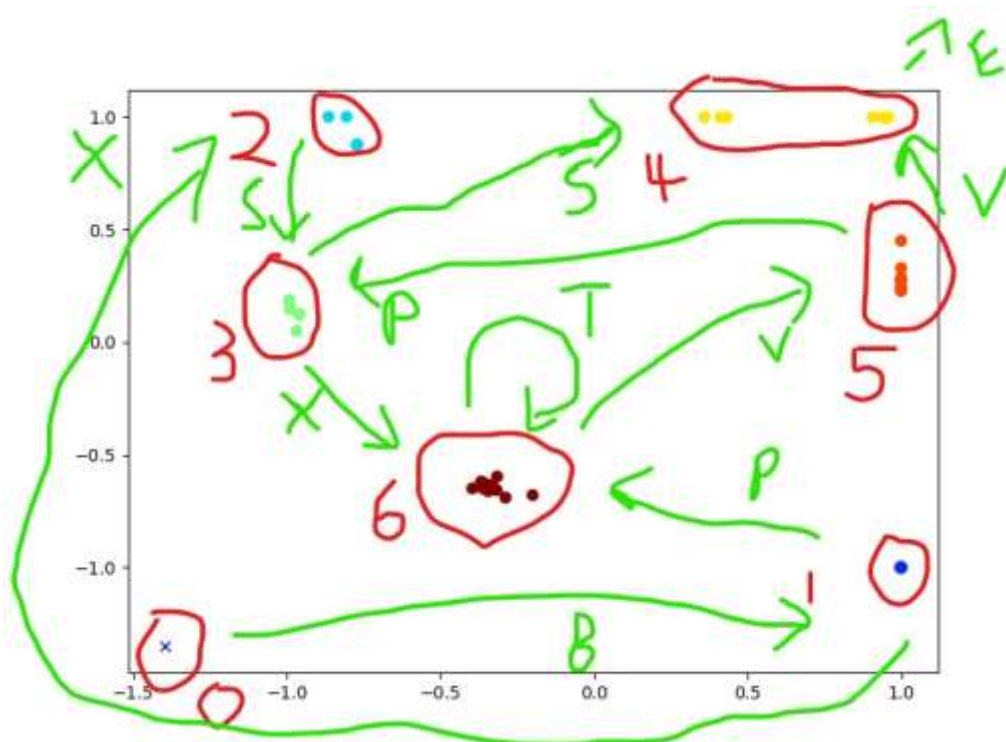
Part 4

Question 1

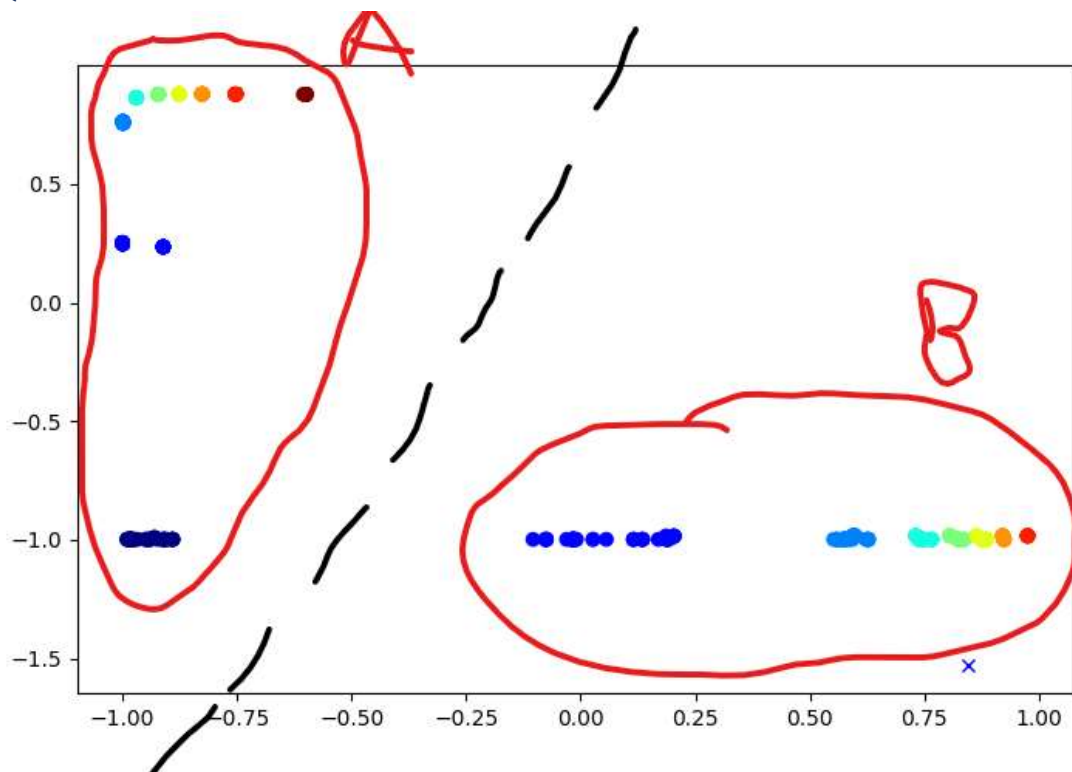
Key for the following diagram –

Red numbers = state

Green Letters = Symbol.



Question 2

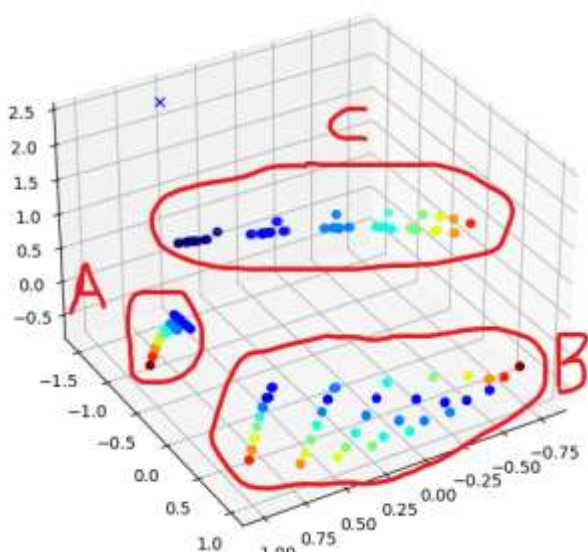


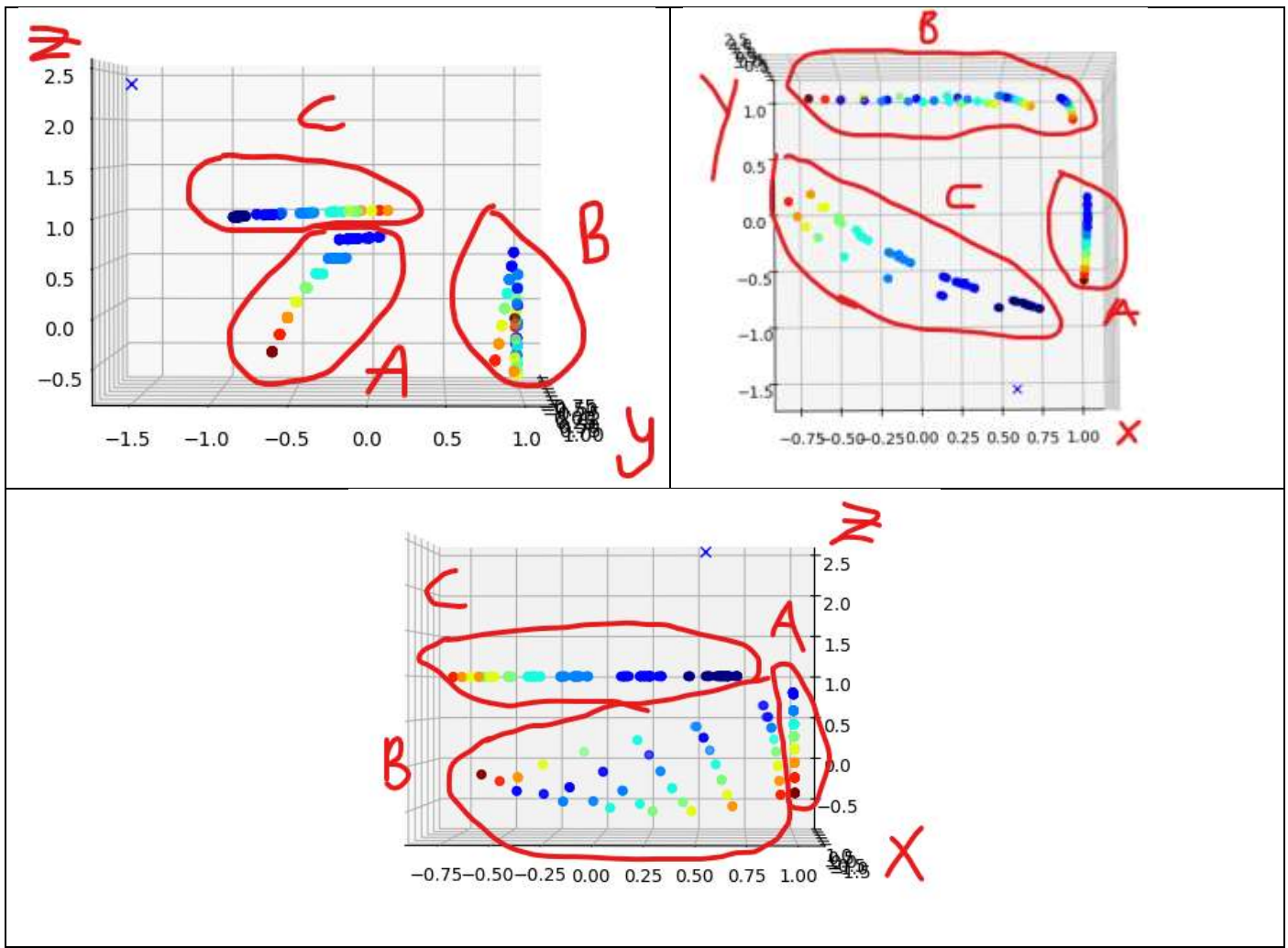
Question 3

There are some letters that the network can predict with certainty since there is a predictable pattern. These are represented by the “redder” dots in Q2. Therefore, the first A of each sequence are “redder” since they are deterministic and have a higher y-axis activation, as shown by the group in the top left-hand corner. All the B’s that occur after the first B are deterministic and are in the bottom right-hand corner since they most activate the hidden unit represented by the x axis.

The darker, blue dots represent the “uncertain” guesses. For example, consecutive A’s cannot continue indefinitely, but there is no deterministic way to determine how many A’s there will be. So, after each “A” guess in a row, the network reduces the probability that the next guess is an A and increases the probability that the next guess is a B

Question 4





Question 5

The 3 hidden units activations correspond to each letter as shown by the graphs above. By isolating a 2d perspective of the 3d image the following is clear –

There is not much x variation in A

There is not much z variation in C

There is not much y variation in B

Of these 3 letters, c has the least variation. C falls almost exactly on the $z = 1$ plane whereas A and B have slight deviations. This is an expected result C is the only class which is purely deterministic. There is some probability involved for A and B since it is uncertain when consecutive As will stop and start a new sequence of Bs. Therefore, the network can predict all the C's, all of the B's except for the first of a sequence, and the first A of the sequence.

Therefore, the hidden node activation of the x axis represents the likelihood of being an A

The hidden node activation of the y axis represents the likelihood of being a B

The hidden node activation of the z axis represents the likelihood of being a C.

Question 6

How the model works –

1. Initialise model and weights
2. For each number in the sequence:
 - a. First make gates based on the weights

- b. Use the sigmoid function to “forget” some context if it is below some threshold
- c. Transform input with sigmoid and tanh
- d. Use a sigmoid function to output result
- e. Update hidden units / gates
- f. Go back to step a.

3. Output result

By plotting the context nodes after every 1000 epochs we can see how the model is “learning”. Early in the training, the context nodes shift wildly since the network is exposed to new information and so may “forget” the required long-term context information. However, after about 20 000 epochs they begin to stabilize.

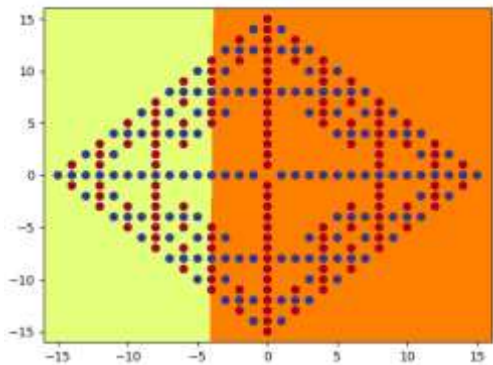
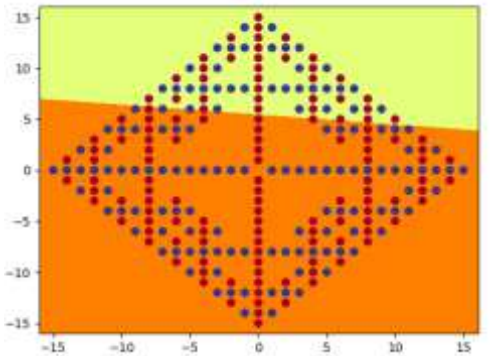
Epoch	Context Node 1	Context Node 2	Context Node 3	Context Node 4
0	-0.5159	0.2825	-0.3055	-0.5741
5 000	-1.3446	-0.9050	-3.0737	-3.8454
10 000	-0.6793	-0.1669	-1.2845	-3.2066
15 000	-1.1163	-0.2603	-2.4591	-5.7861
20 000	1.1615	-3.2945	0.3195	0.6405
25 000	0.9829	-2.6177	0.3974	0.3096
30 000	0.9870	-2.2228	0.9371	-0.6356
35 000	0.7659	-2.2290	0.8086	-0.3102
40 000	0.9626	-1.3299	0.8989	-0.4284
45 000	0.9744	-2.2622	1.0170	-1.1428
50 000	0.9464	-1.9169	0.9432	-0.8253

Therefore, over time the LSTM can effectively have “long term memory” through these context units as they approach a value approximately [1, -2, 1, -1] which will lead to the highest accuracy in its predictions.

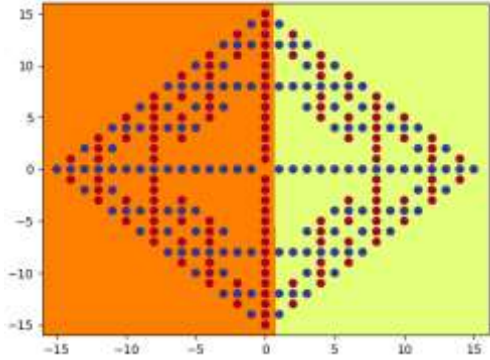
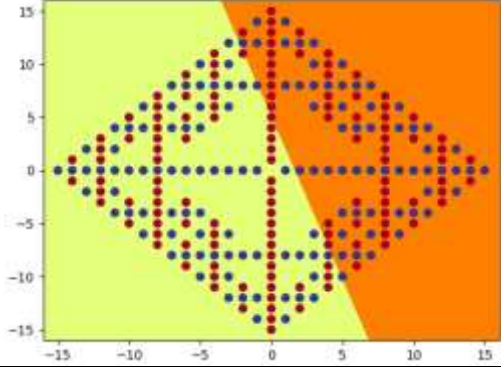
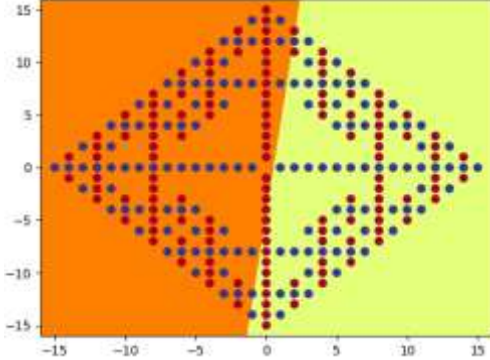
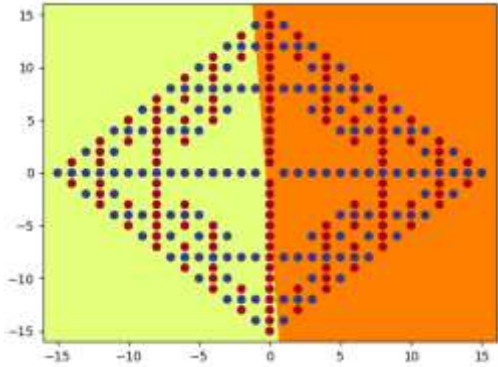
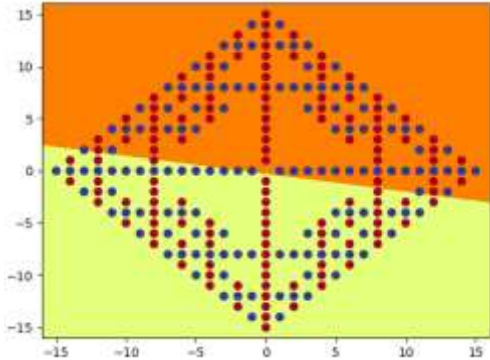
Appendices

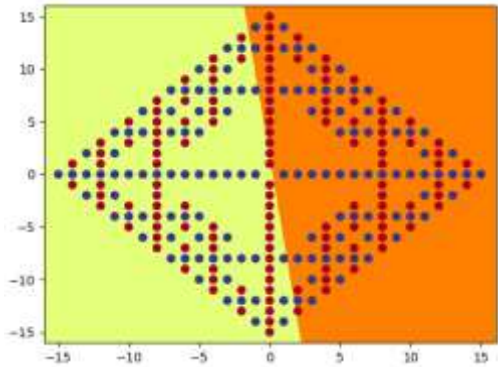
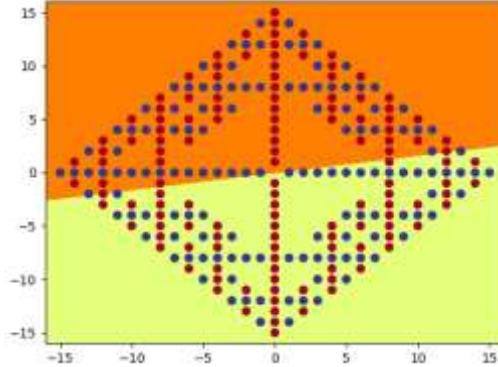
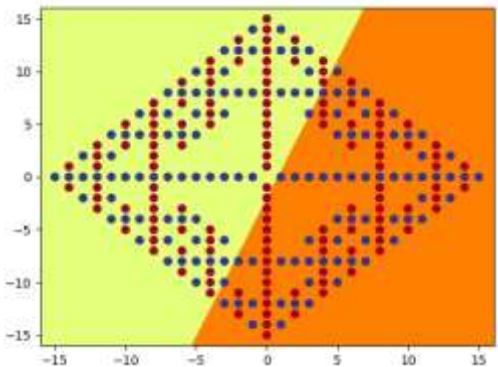
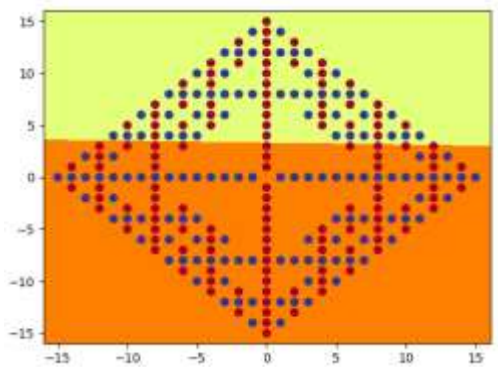
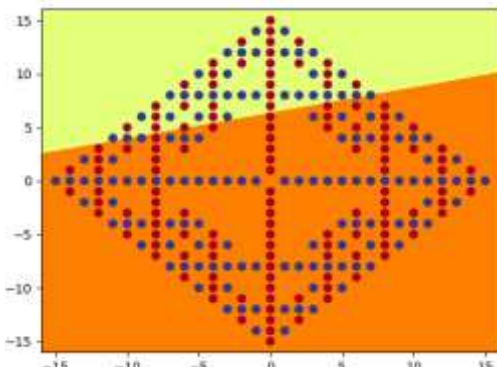
Appendix A

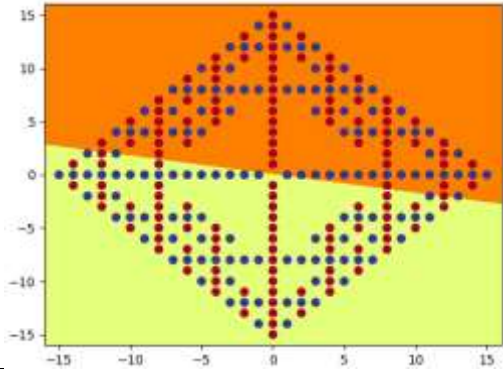
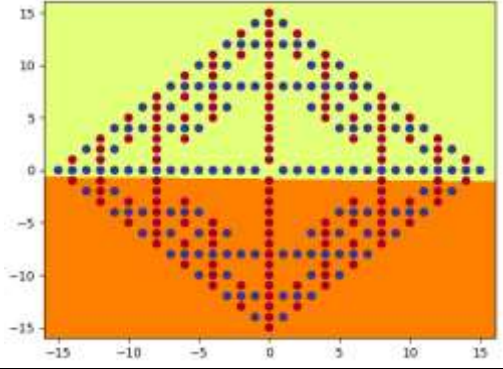
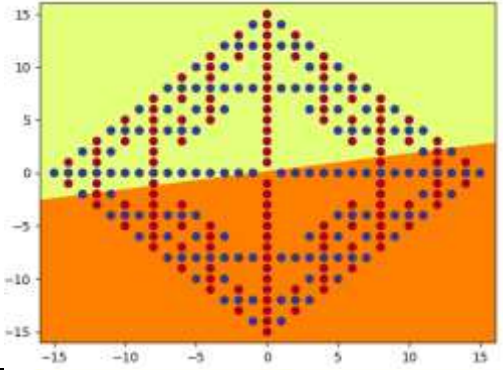
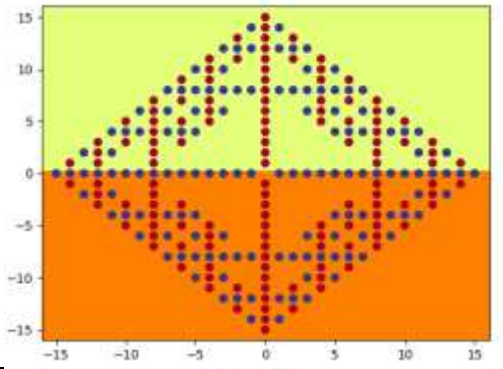
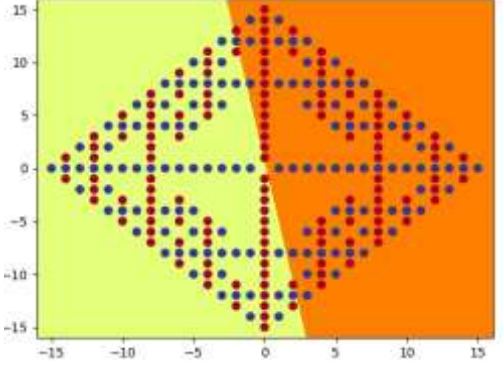
Hidden unit plots for the 3-layer full network–

Node number	Hidden layer	Plot
0	1	
1	1	

2	1	
3	1	
4	1	
5	1	
6	1	

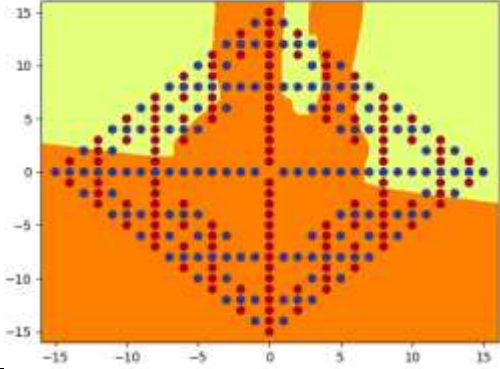
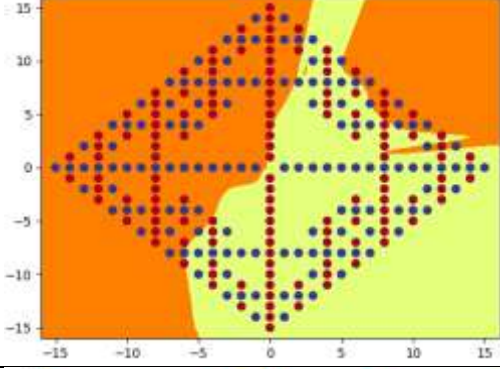
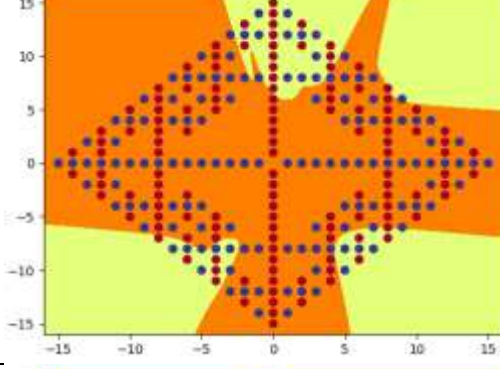
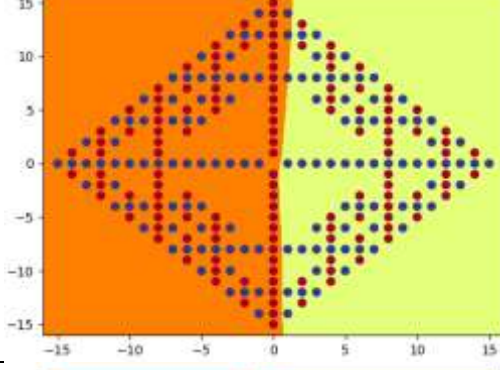
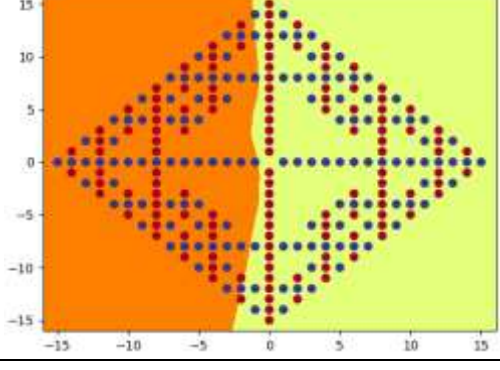
7	1		
8	1		
9	1		
10	1		
11	1		

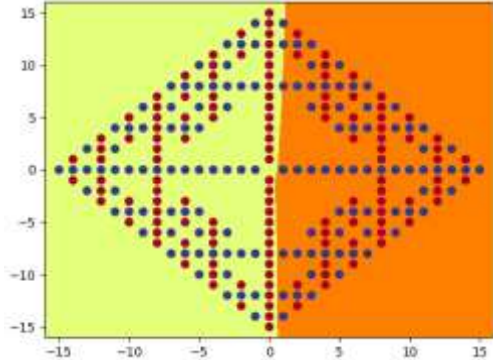
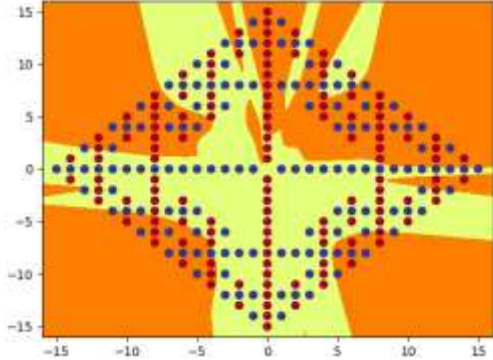
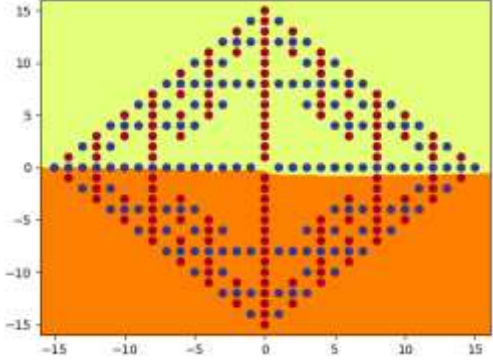
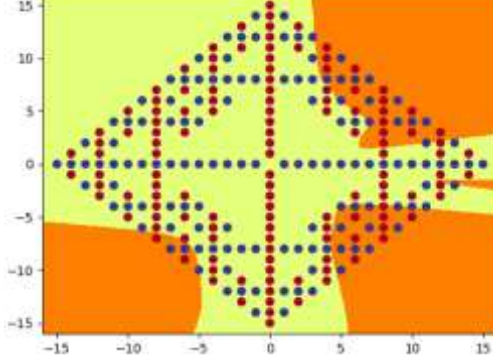
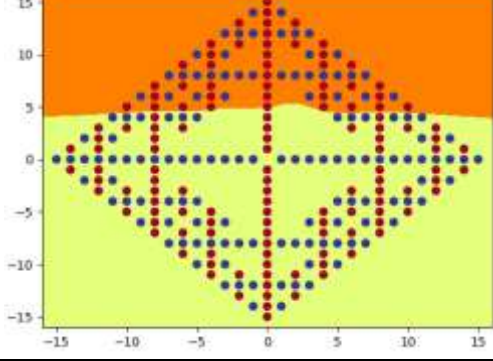
12	1	
13	1	
14	1	
15	1	
16	1	

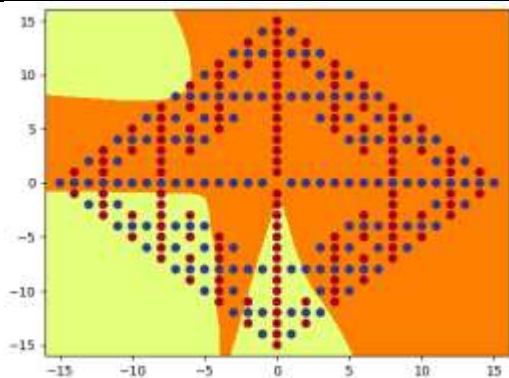
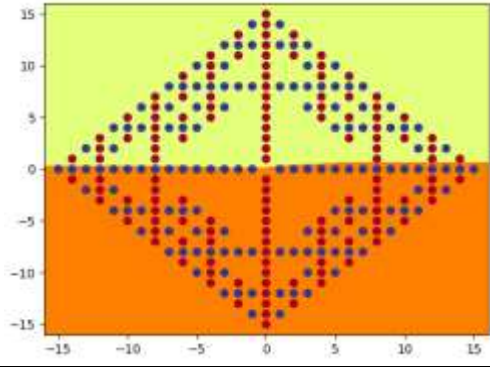
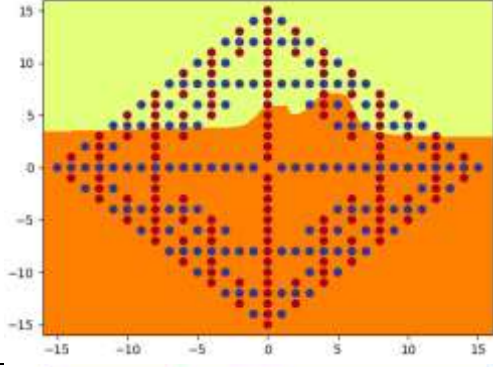
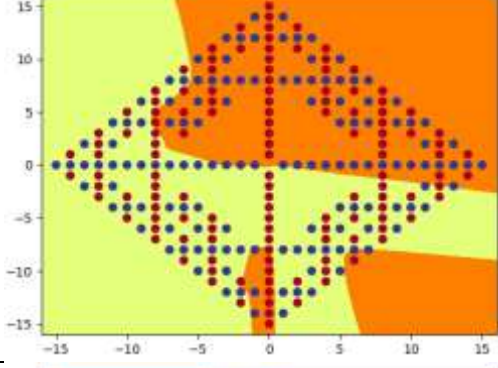
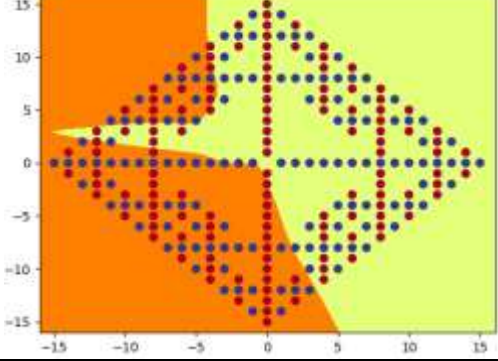
17	1	
18	1	
19	1	
20	1	
21	1	

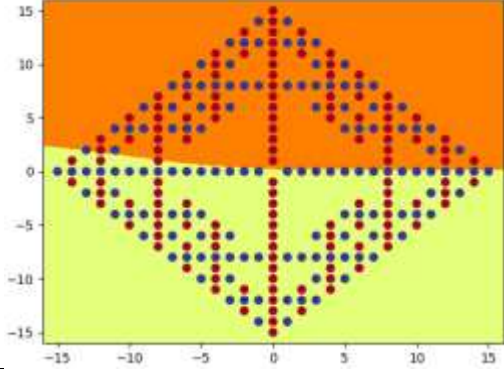
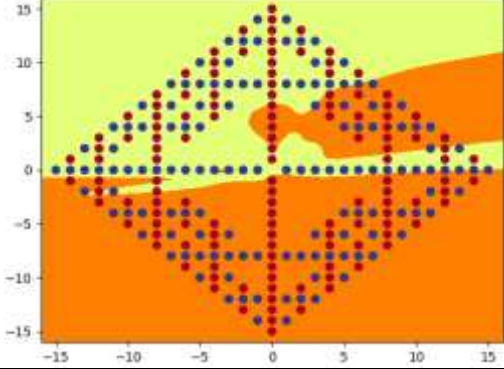
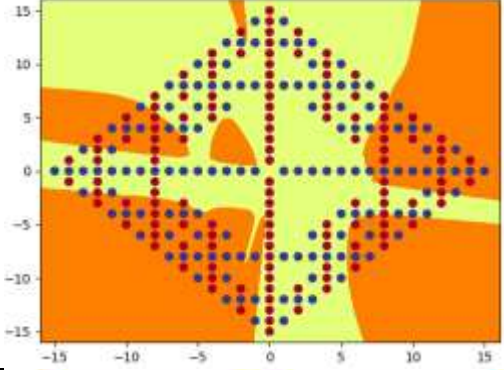
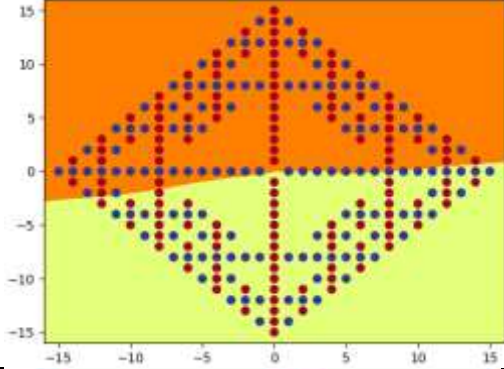
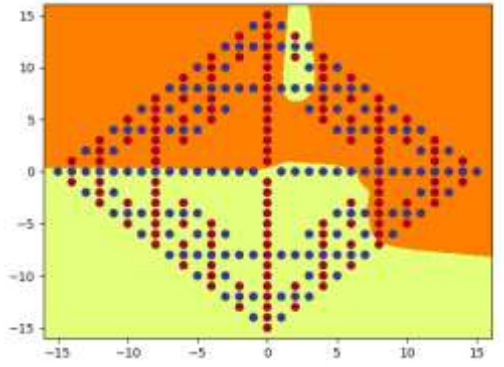
22	1	
23	1	
24	1	
25	1	
26	1	

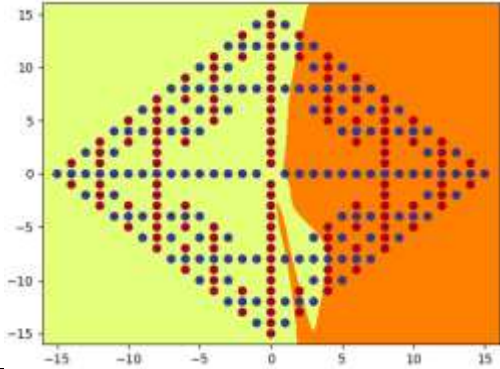
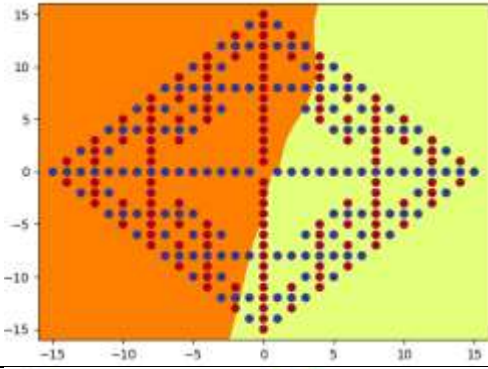
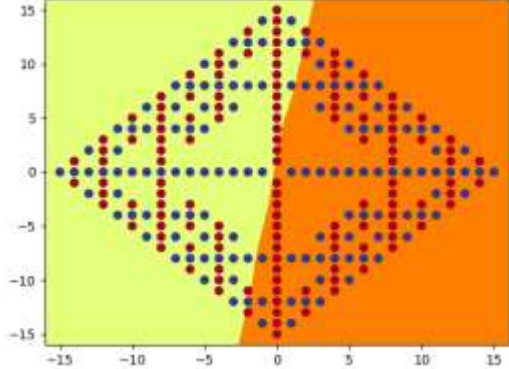
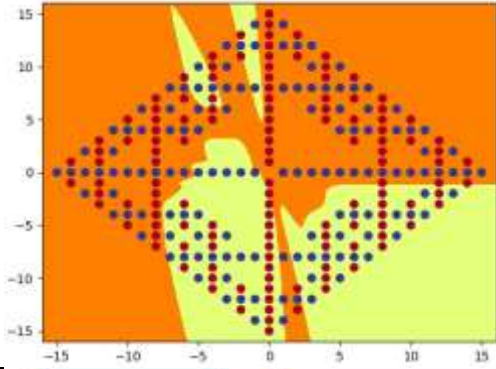
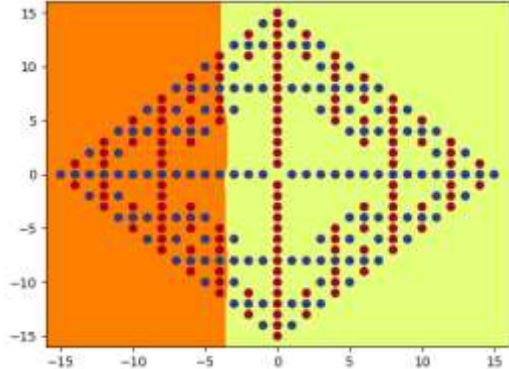
27	1		
28	1		
29	1		
0	2		
1	2		

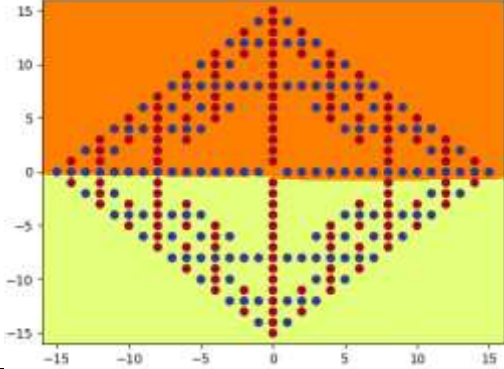
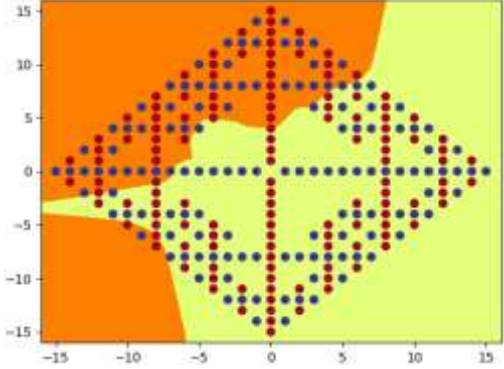
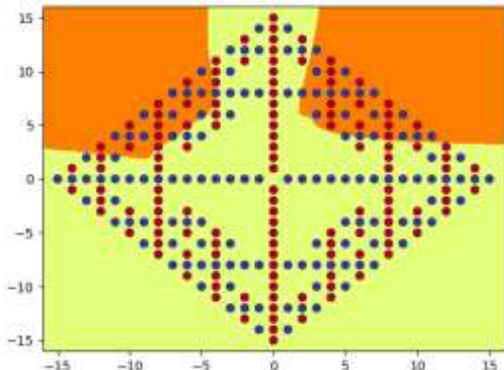
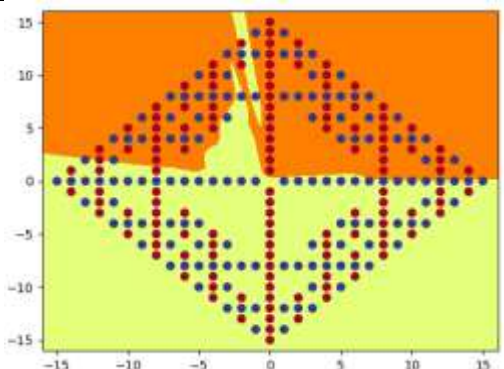
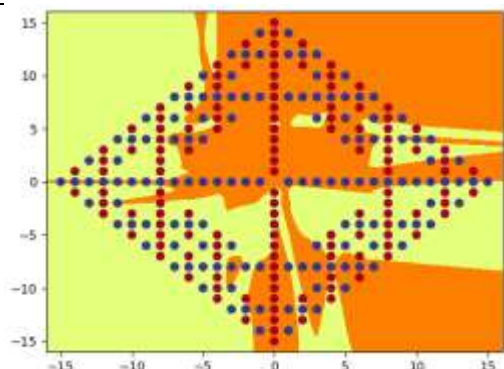
2	2		
3	2		
4	2		
5	2		
6	2		

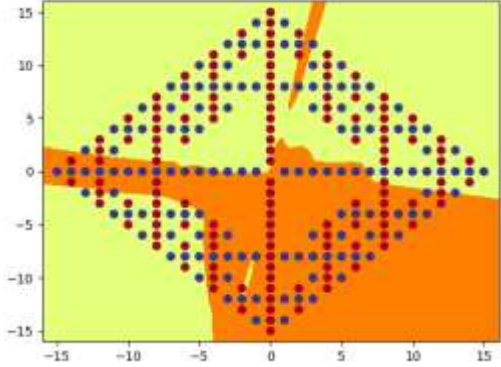
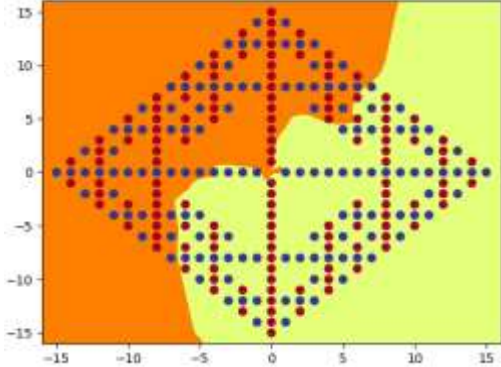
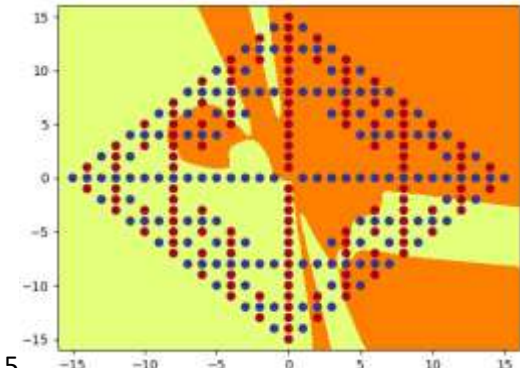
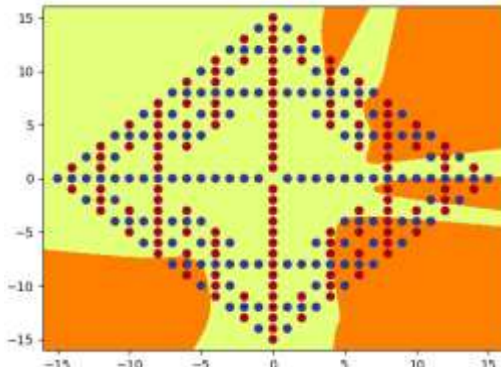
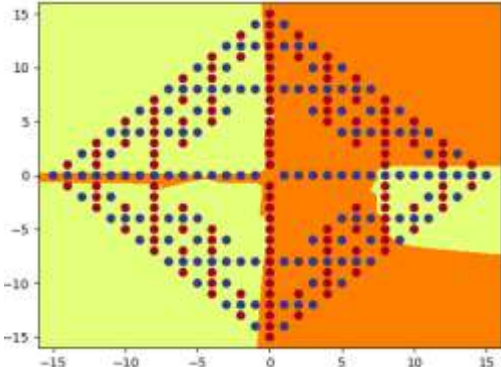
7	2		
8	2		
9	2		
10	2		
11	2		

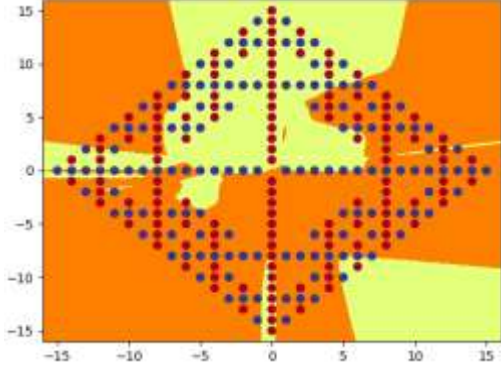
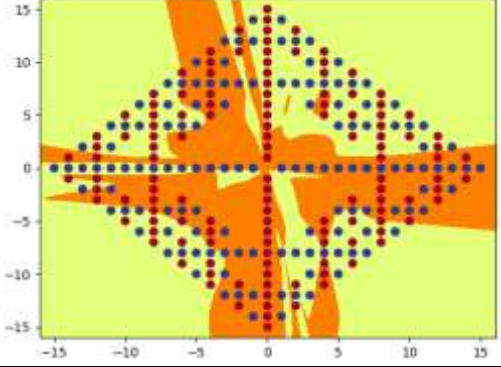
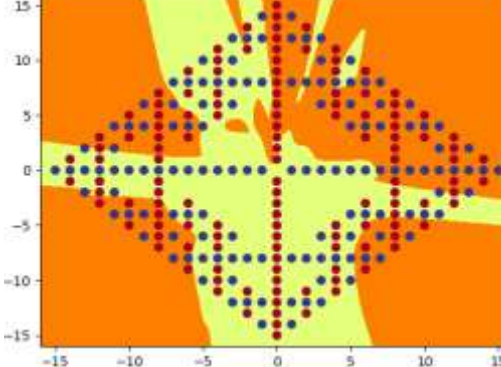
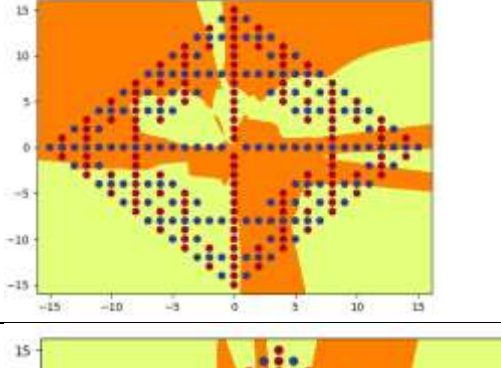
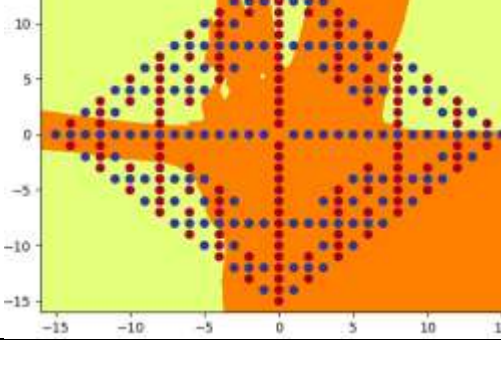
12	2	
13	2	
14	2	
15	2	
16	2	

17	2		
18	2		
19	2		
20	2		
21	2		

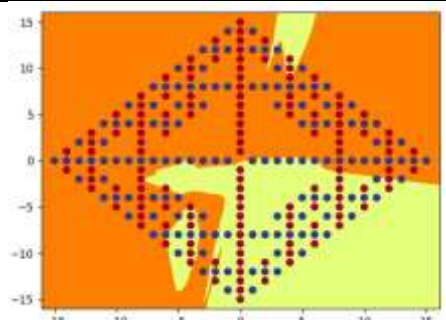
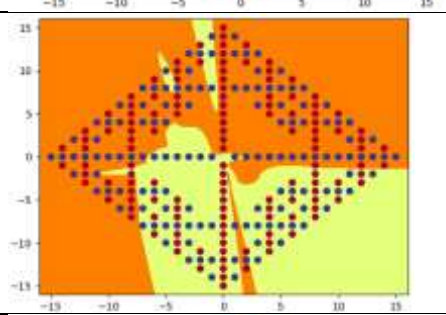
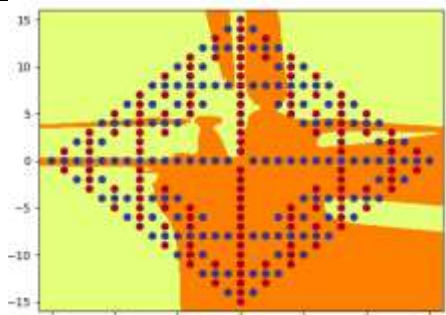
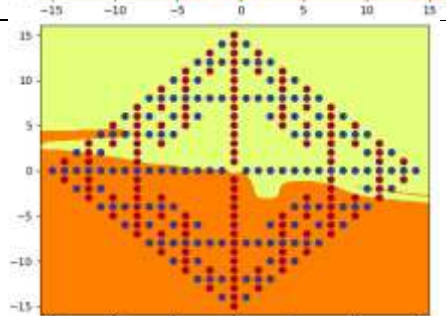
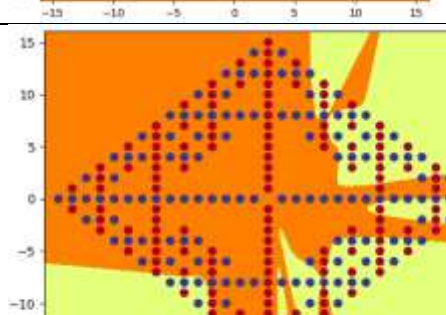
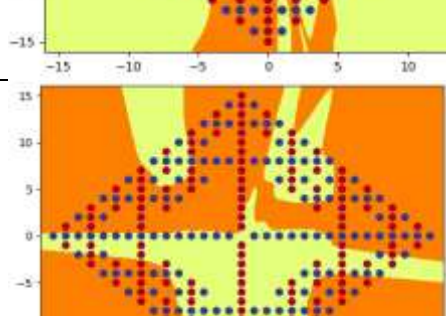
22	2	
23	2	
24	2	
25	2	
26	2	

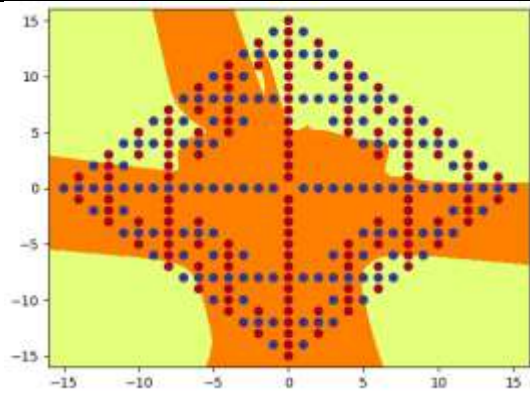
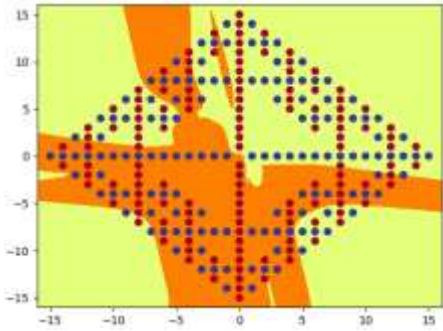
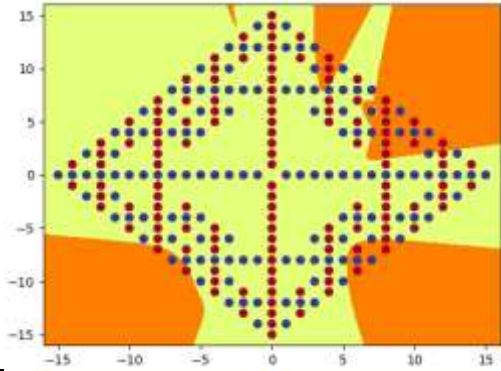
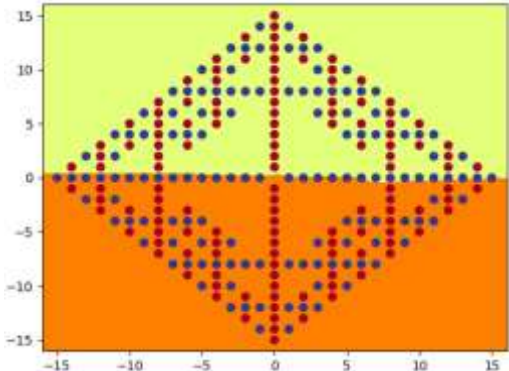
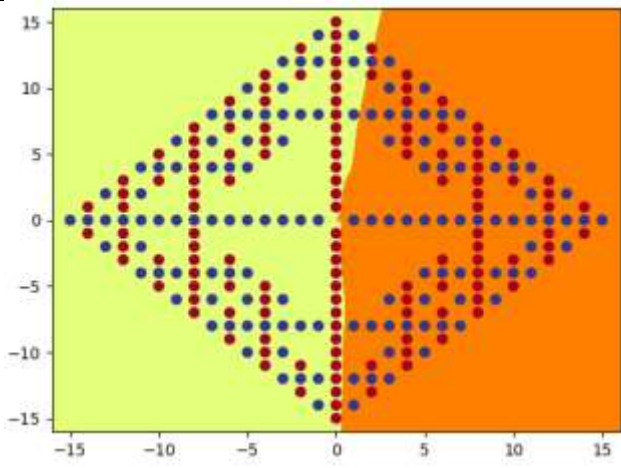
27	2		
28	2		
29	2		
0	3		
1	3		

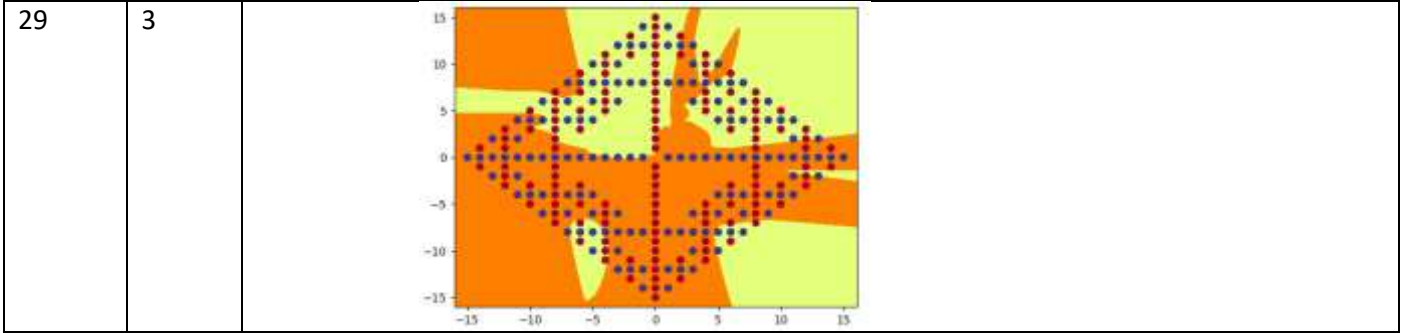
2	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. Red points form a central cross and four diagonal arms. Blue points form a grid-like pattern. The background is divided into orange and yellow regions by a complex, non-linear boundary.</p>
3	3	 <p>A scatter plot similar to the first, but with a different distribution of orange and yellow background regions, indicating a different model or parameter setting.</p>
4	3	 <p>A scatter plot with a more complex and fragmented background pattern of orange and yellow regions compared to the previous plots.</p>
5	3	 <p>A scatter plot showing a different configuration of background regions, with a more pronounced central yellow area.</p>
6	3	 <p>A scatter plot with a background pattern that is more similar to the first plot but with slight variations in the orange and yellow regions.</p>

7	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a complex pattern of orange and yellow regions. Red and blue points are distributed in a diamond shape centered at the origin, with points forming a grid-like pattern within the diamond.</p>
8	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a complex pattern of orange and yellow regions. Red and blue points are distributed in a diamond shape centered at the origin, with points forming a grid-like pattern within the diamond.</p>
9	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a complex pattern of orange and yellow regions. Red and blue points are distributed in a diamond shape centered at the origin, with points forming a grid-like pattern within the diamond.</p>
10	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a complex pattern of orange and yellow regions. Red and blue points are distributed in a diamond shape centered at the origin, with points forming a grid-like pattern within the diamond.</p>
11	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a complex pattern of orange and yellow regions. Red and blue points are distributed in a diamond shape centered at the origin, with points forming a grid-like pattern within the diamond.</p>

12	3	
13	3	
14	3	
15	3	
16	3	
17	3	

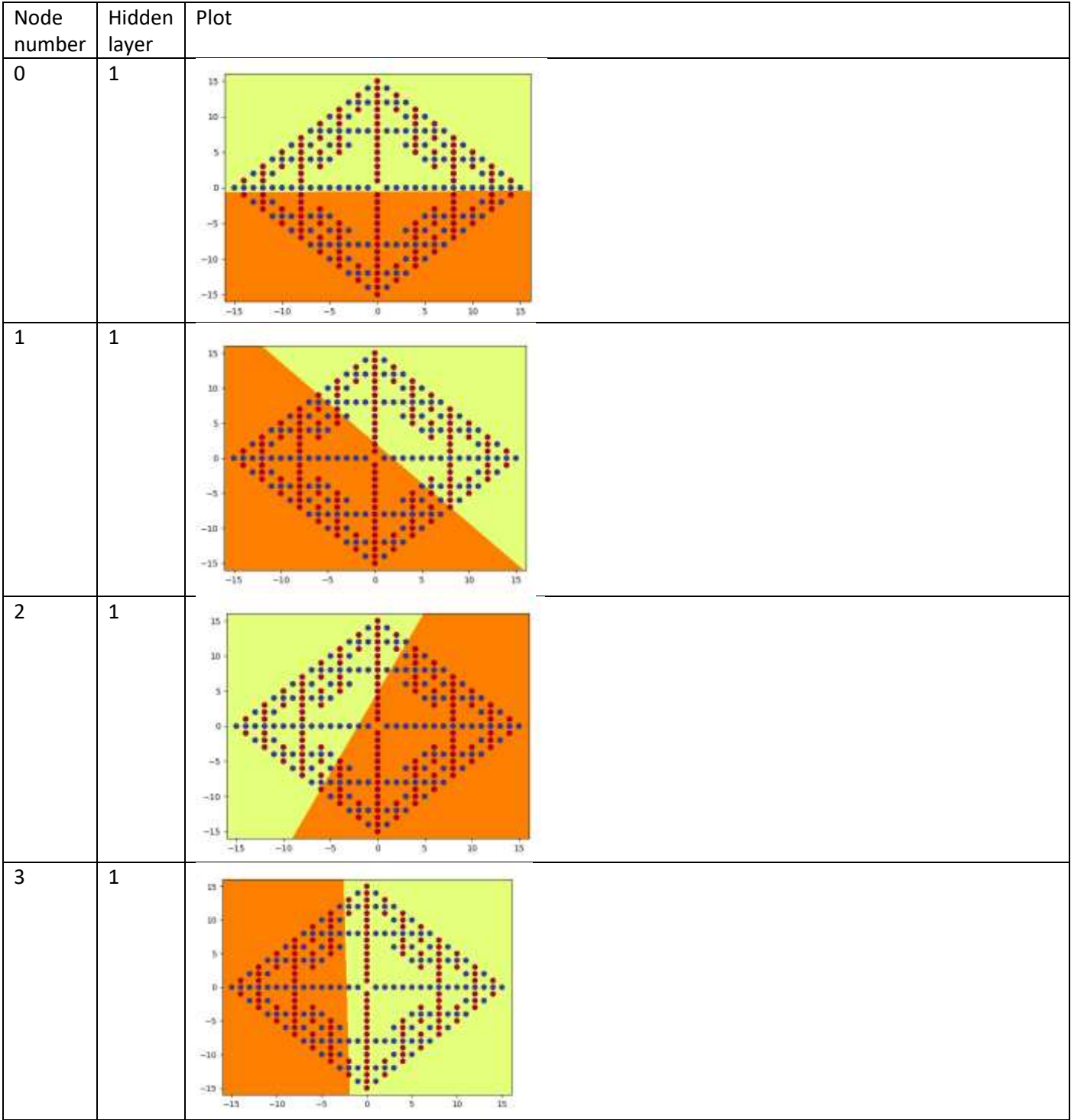
18	3	
19	3	
20	3	
21	3	
22	3	
23	3	

24	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a yellow and orange checkerboard pattern. Red and blue points are distributed in a diamond shape centered at (0,0), with red points forming a grid-like pattern and blue points filling the surrounding space within the diamond.</p>
25	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a yellow and orange checkerboard pattern. Red and blue points are distributed in a diamond shape centered at (0,0), with red points forming a grid-like pattern and blue points filling the surrounding space within the diamond.</p>
26	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a yellow and orange checkerboard pattern. Red and blue points are distributed in a diamond shape centered at (0,0), with red points forming a grid-like pattern and blue points filling the surrounding space within the diamond.</p>
27	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a yellow and orange checkerboard pattern. Red and blue points are distributed in a diamond shape centered at (0,0), with red points forming a grid-like pattern and blue points filling the surrounding space within the diamond.</p>
28	3	 <p>A scatter plot on a coordinate system from -15 to 15 on both axes. The background is a yellow and orange checkerboard pattern. Red and blue points are distributed in a diamond shape centered at (0,0), with red points forming a grid-like pattern and blue points filling the surrounding space within the diamond.</p>



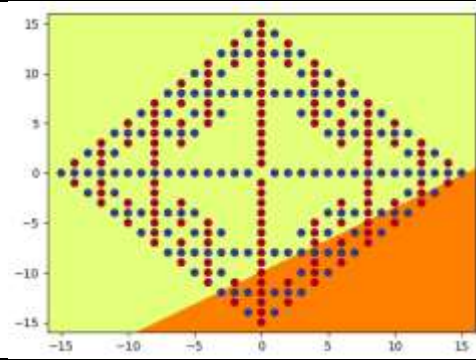
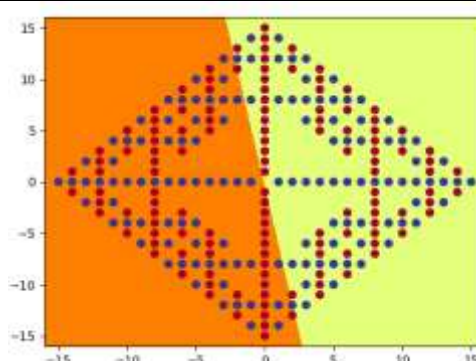
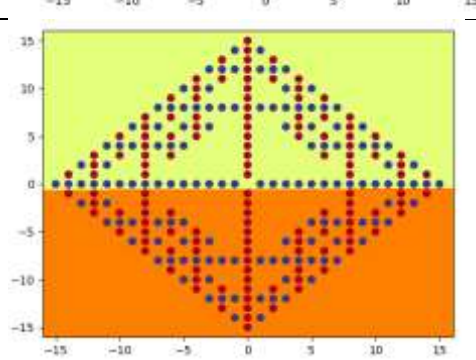
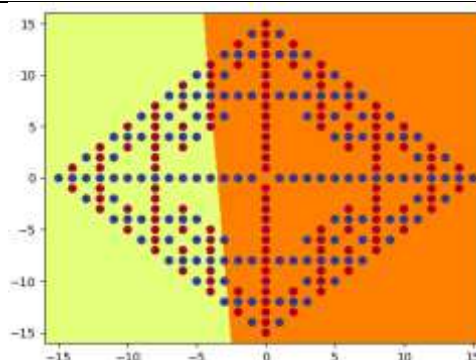
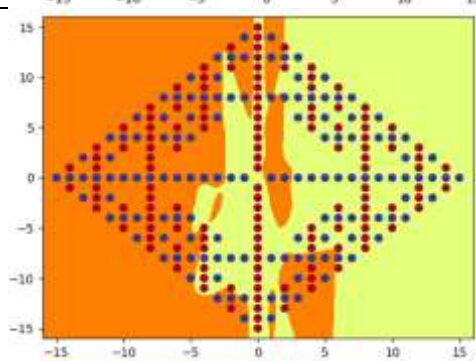
Appendix B

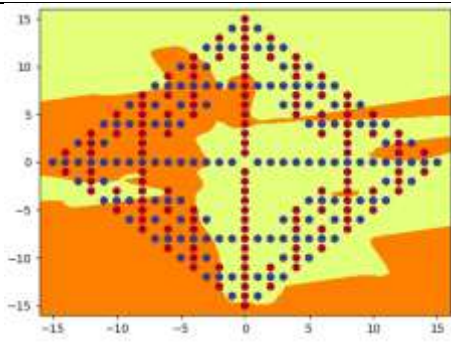
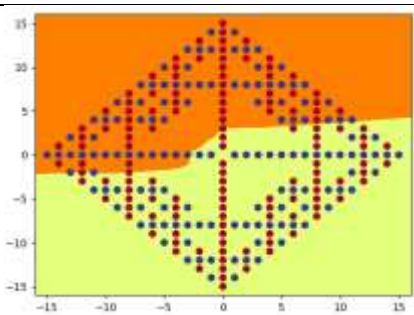
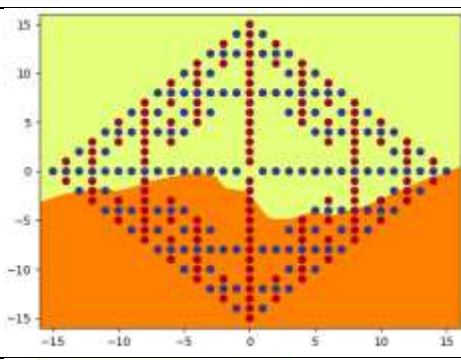
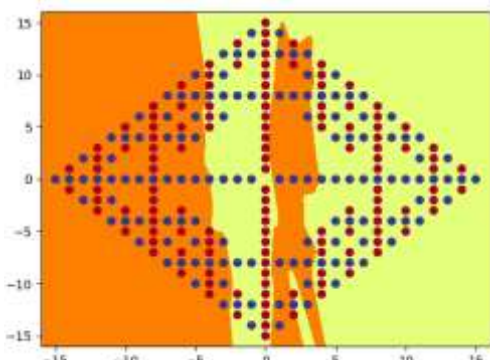
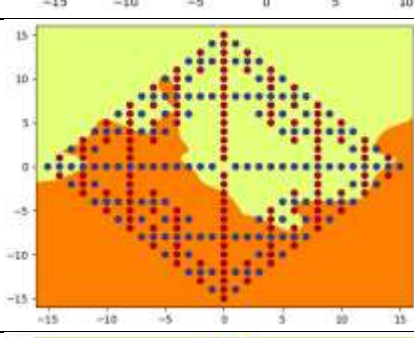
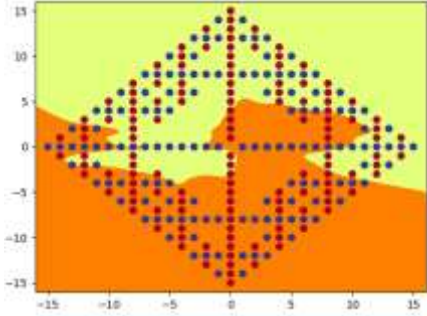
Hidden unit plots for the dense network

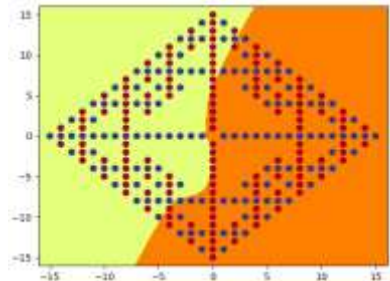
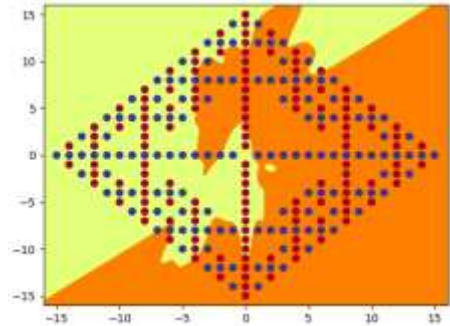
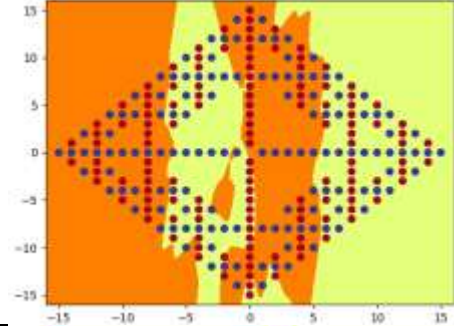
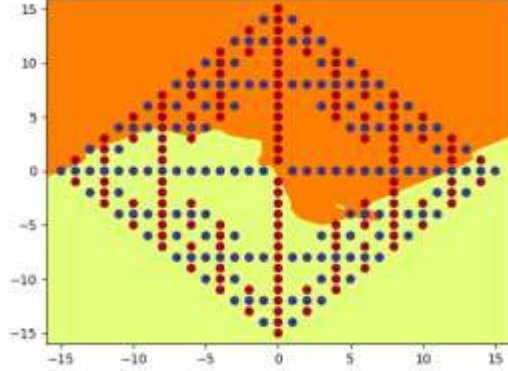
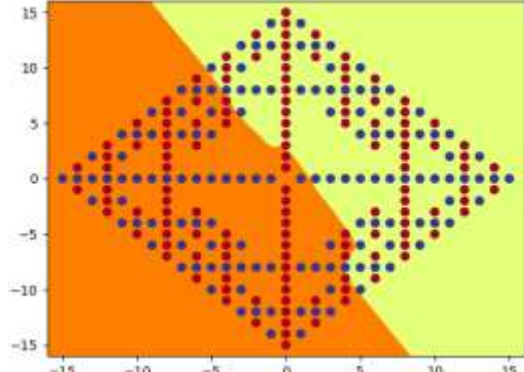


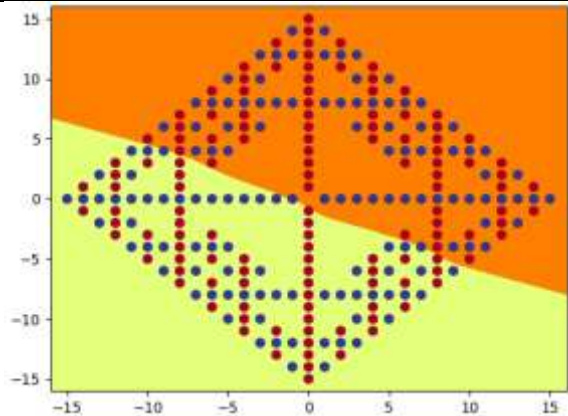
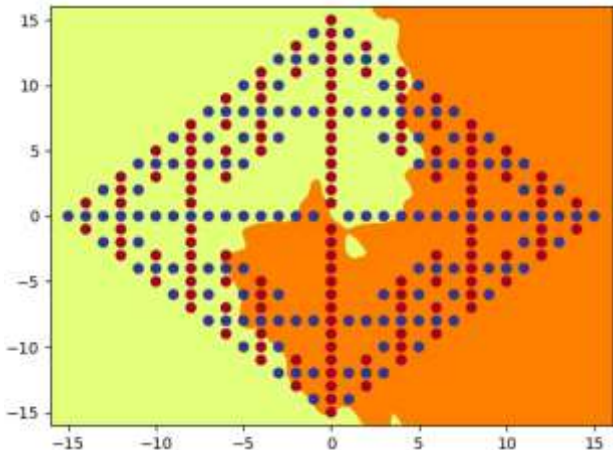
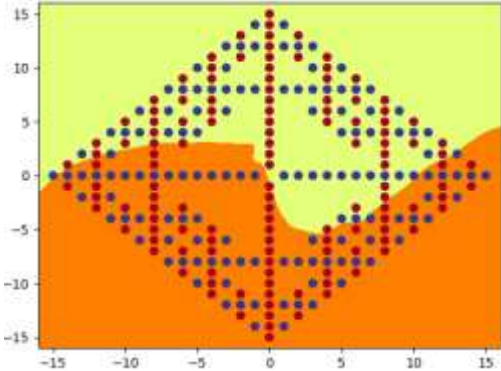
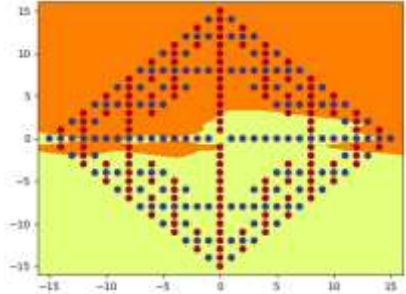
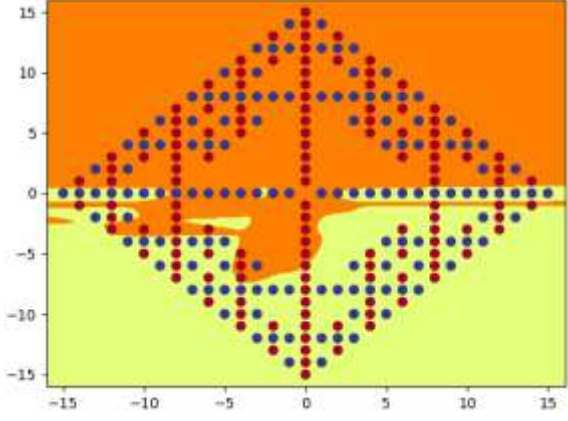
4	1	
5	1	
6	1	
7	1	
8	1	
9	1	

10	1	
11	1	
12	1	
13	1	
14	1	
15	1	

16	1	
17	1	
18	1	
19	1	
0	2	

1	2	
2	2	
3	2	
4	2	
5	2	
6	2	

7	2	 A scatter plot on a coordinate plane with x and y axes ranging from -15 to 15. The background is divided into yellow and orange regions by a complex, non-linear boundary. Data points are plotted in a grid-like pattern, with red dots concentrated in the orange regions and blue dots in the yellow regions.
8	2	 A scatter plot on a coordinate plane with x and y axes ranging from -15 to 15. The background is divided into yellow and orange regions by a complex, non-linear boundary. Data points are plotted in a grid-like pattern, with red dots concentrated in the orange regions and blue dots in the yellow regions.
9	2	 A scatter plot on a coordinate plane with x and y axes ranging from -15 to 15. The background is divided into yellow and orange regions by a complex, non-linear boundary. Data points are plotted in a grid-like pattern, with red dots concentrated in the orange regions and blue dots in the yellow regions.
10	2	 A scatter plot on a coordinate plane with x and y axes ranging from -15 to 15. The background is divided into yellow and orange regions by a complex, non-linear boundary. Data points are plotted in a grid-like pattern, with red dots concentrated in the orange regions and blue dots in the yellow regions.
11	2	 A scatter plot on a coordinate plane with x and y axes ranging from -15 to 15. The background is divided into yellow and orange regions by a complex, non-linear boundary. Data points are plotted in a grid-like pattern, with red dots concentrated in the orange regions and blue dots in the yellow regions.

12	2	
13	2	
14	2	
15	2	
16	2	

17	2	
18	2	
19	2	