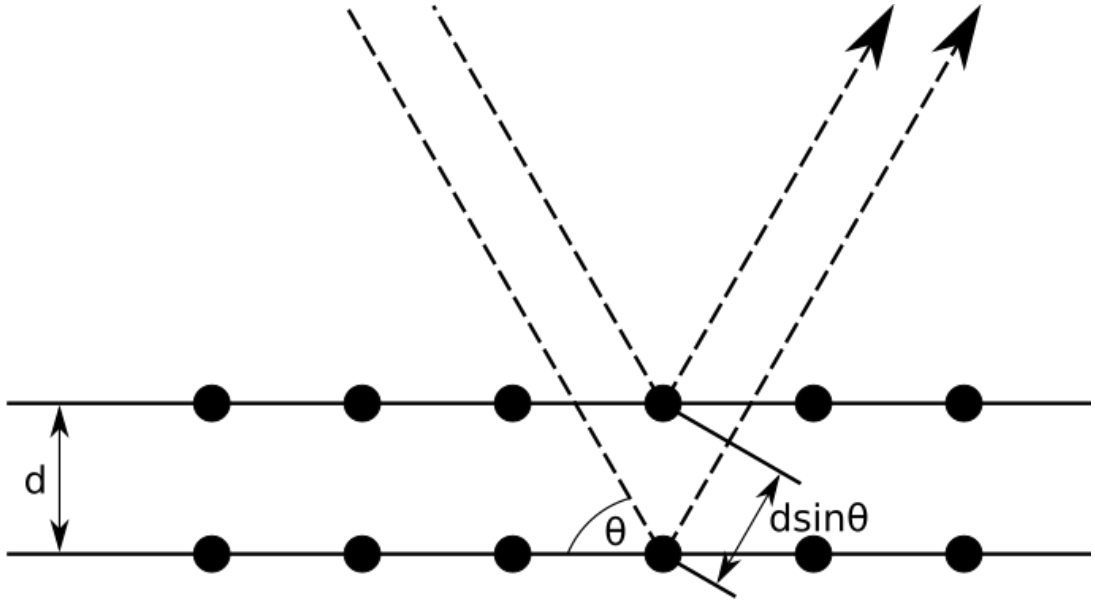
Utilising Machine Learning on XRD Crystal Classification

侯奕安、吳政蔚、蔡杰達

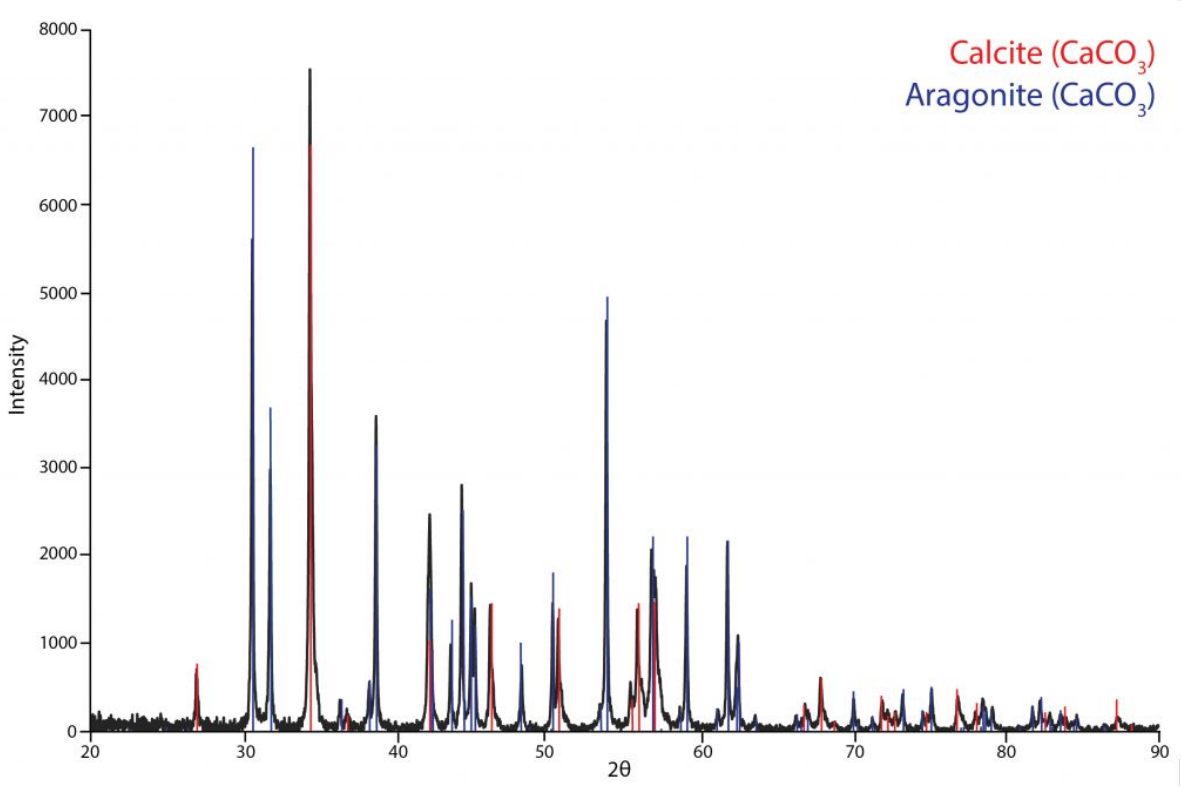
What is XRD ?

Bragg's Law

$$2dsin\theta = n\lambda$$



XRD Spectrum



Miller Indices

Consider a lattice plane with normal vector:

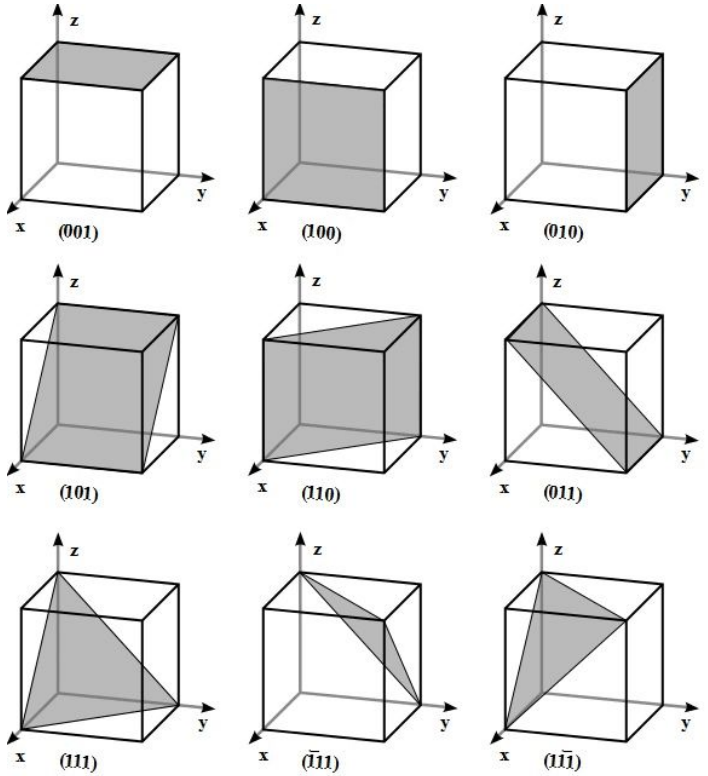
(1/h, 1/k, 1/l)

Then the Miller indices are denoted as:

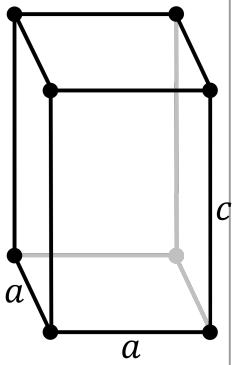
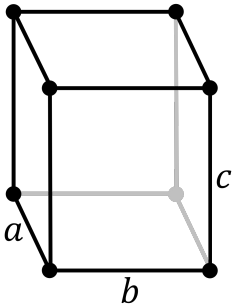
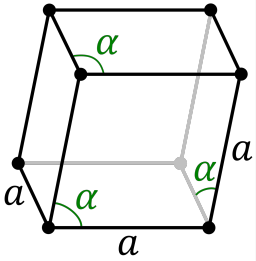
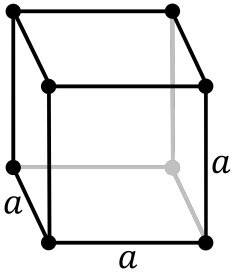
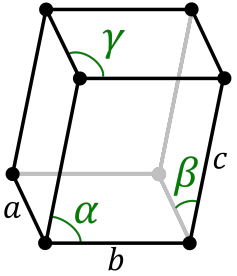
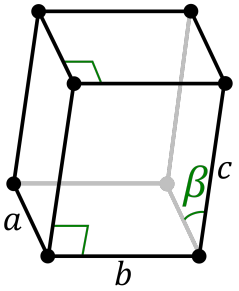
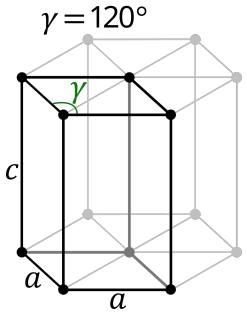
(h, k, l)

Classes are determined by relations between (h, k, l) and spacing between adjacent lattice plane.

Eg. cubic, hexagonal, etc

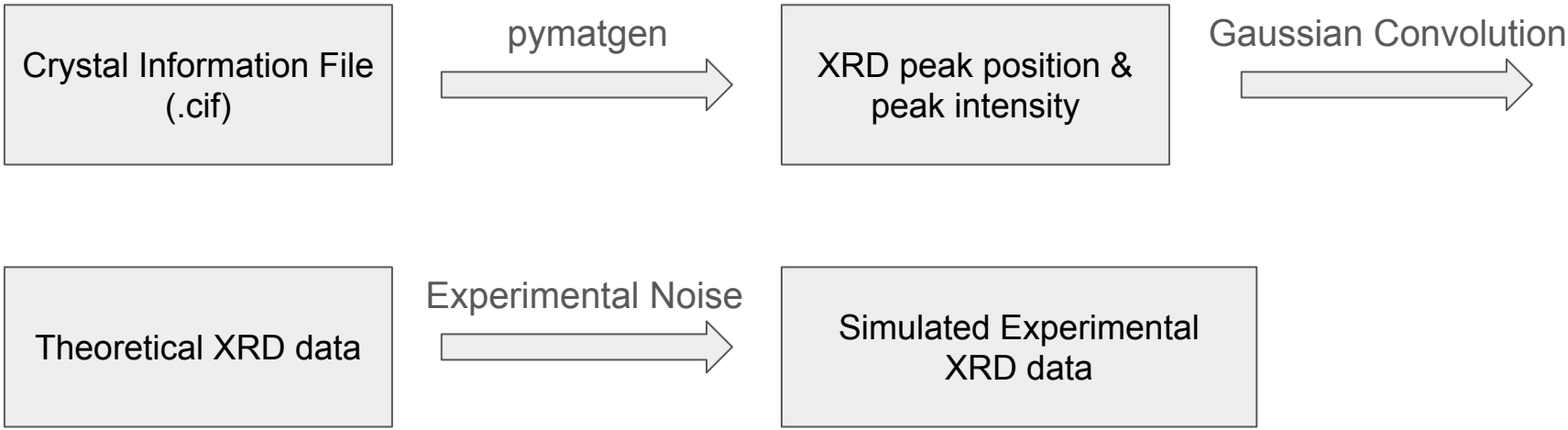


Lattices

Tetragonal(0)	orthorhombic (1)	trigonal(2)	cubic(3)	triclinic(4)	monoclinic(5)	hexagonal(6)
						

Data Preparation

Data preparation

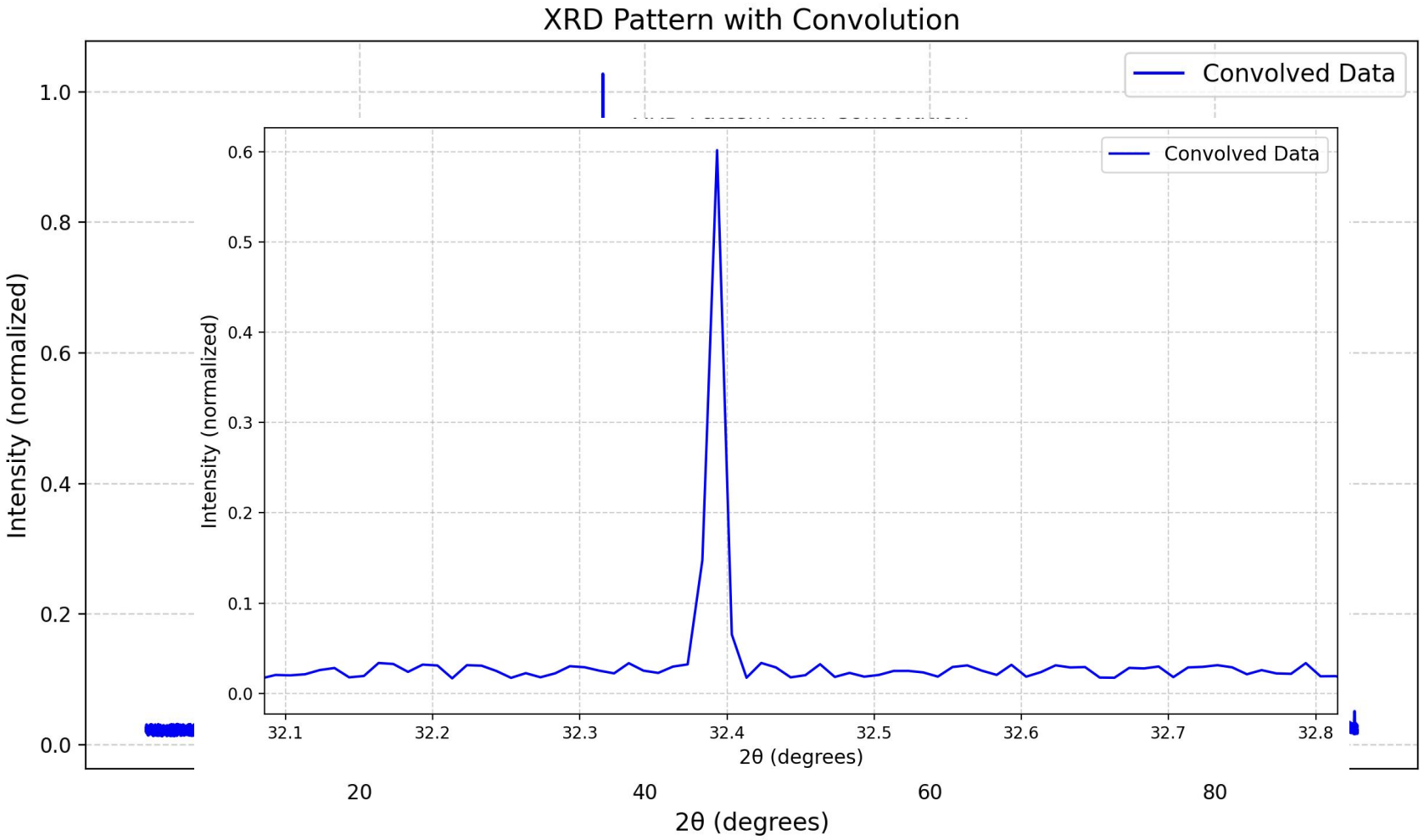


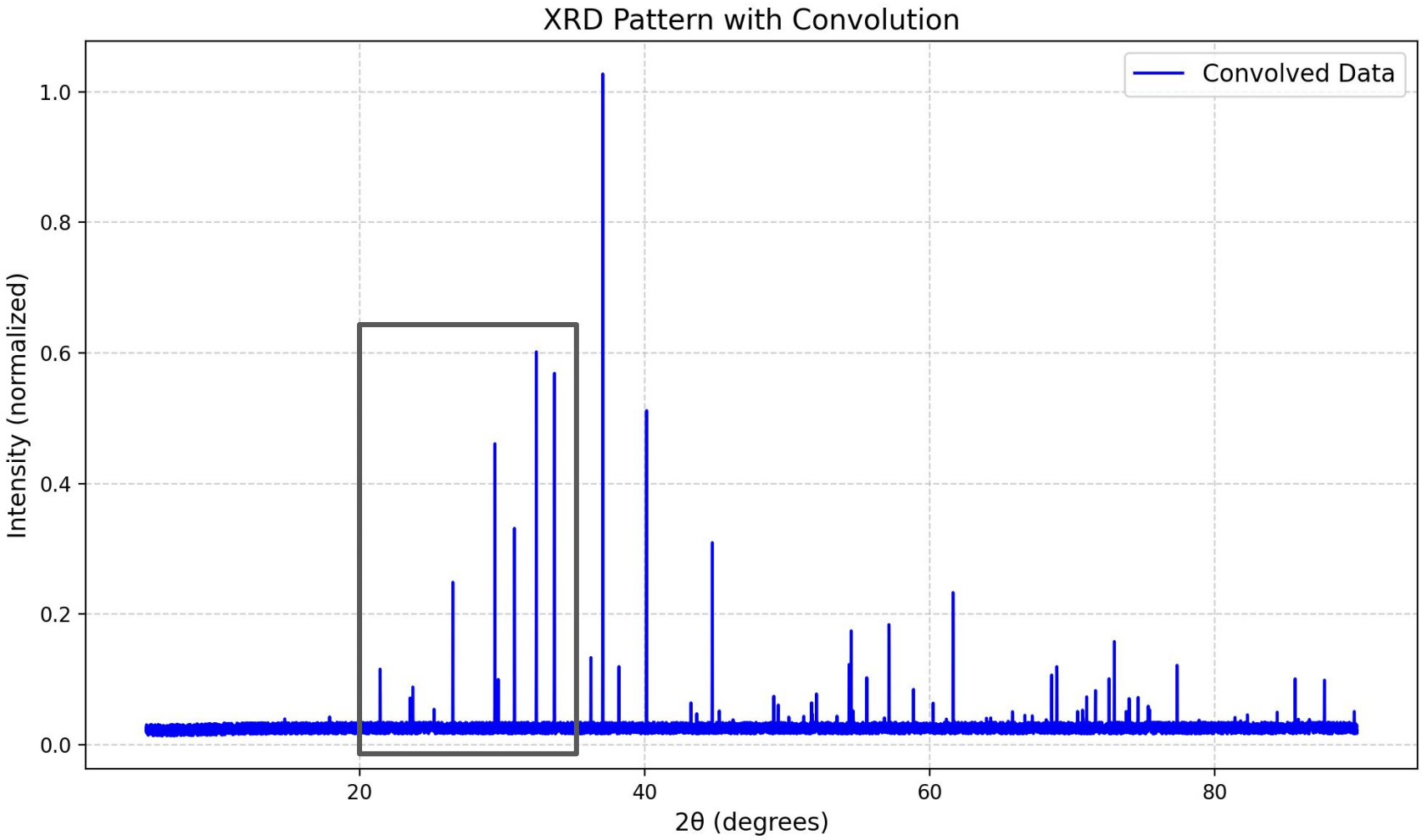
Data preparation (CIF)

```
_chemical_formula_sum              'As Gd Te'
_chemical_formula_weight           359.77
_chemical_name_systematic
;
Gadolinium Asernic Telluride
;
_space_group_IT_number             62
_symmetry_cell_setting             orthorhombic
_symmetry_space_group_name_Hall    '-P 2ac 2n'
_symmetry_space_group_name_H-M     'P n m a'
_atom_sites_solution_hydrogens     geom
_atom_sites_solution_primary       direct
_atom_sites_solution_secondary     difmap
_audit_creation_method             SHELXL-97
_cell_angle_alpha                  90.00
_cell_angle_beta                   90.00
_cell_angle_gamma                  90.00
_cell_formula_units_Z              4
_cell_length_a                     7.5611(15)
_cell_length_b                     4.0510(8)
_cell_length_c                     9.920(2)
_cell_measurement_temperature      153(2)
_cell_volume                       303.85(10)
```

- We remove file with large crystal size (50 Å) and organic substance
- There are 19,525 data in total

Da


XRD Pattern with Convolution
Convolved Data
Intensity (normalized)
2θ (degrees)
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
Convolved Data
Intensity (normalized)
2θ (degrees)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
32.1
32.2
32.3
32.4
32.5
32.6
32.7
32.8

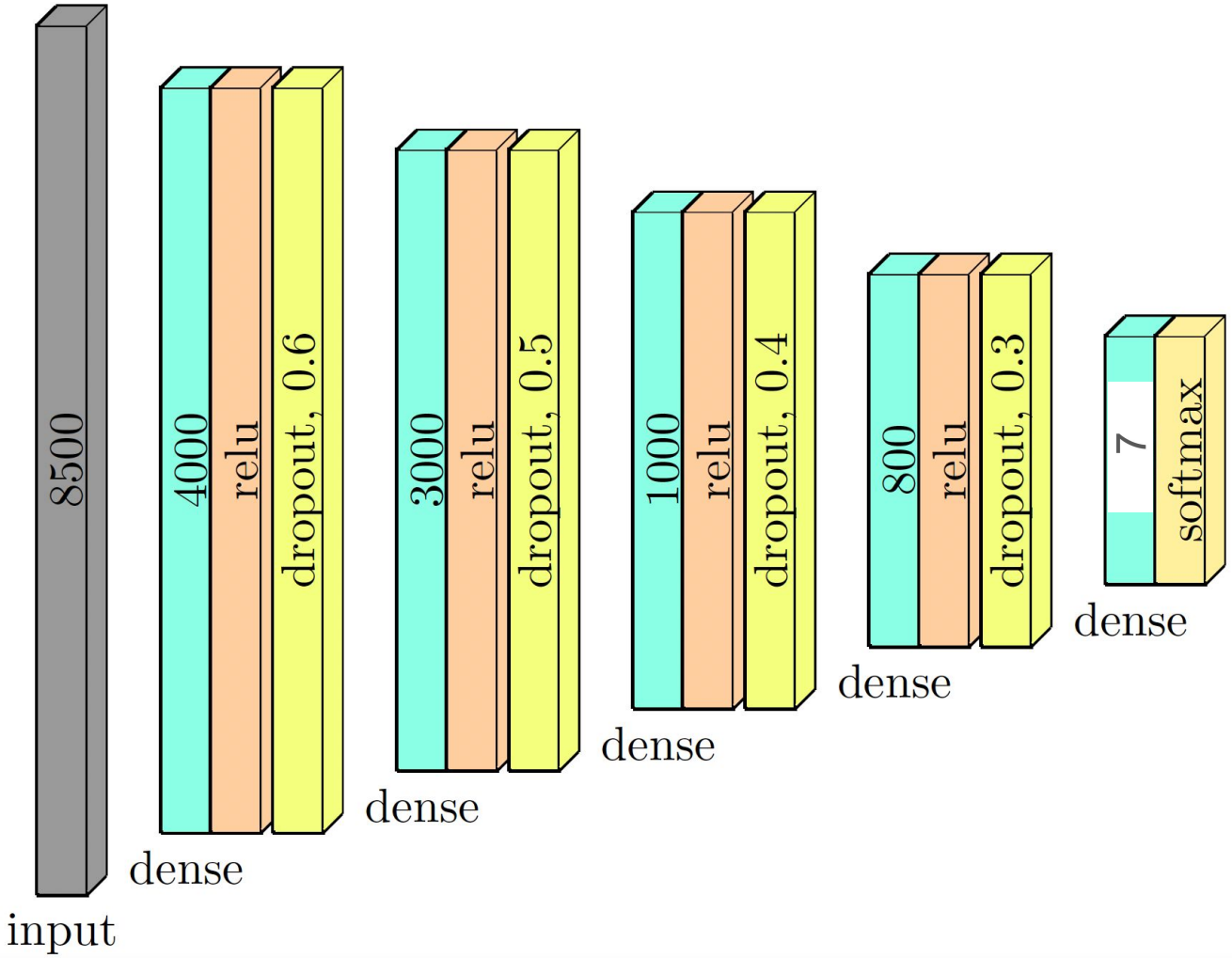

XRD Pattern with Convolution
Convolved Data
Intensity (normalized)
2θ (degrees)
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80

Data train/test split

```
# Split data based on split_info
split_info = {
    "tetragonal": (3000, 500, 100),
    "orthorhombic": (3500, 500, 300),
    "trigonal": (2000, 500, 100),
    "cubic": (2000, 500, 100),
    "triclinic": (1800, 500, 10),
    "monoclinic": (3200, 500, 300),
    "hexagonal": (100, 10, 5)
}
```

1. Dense

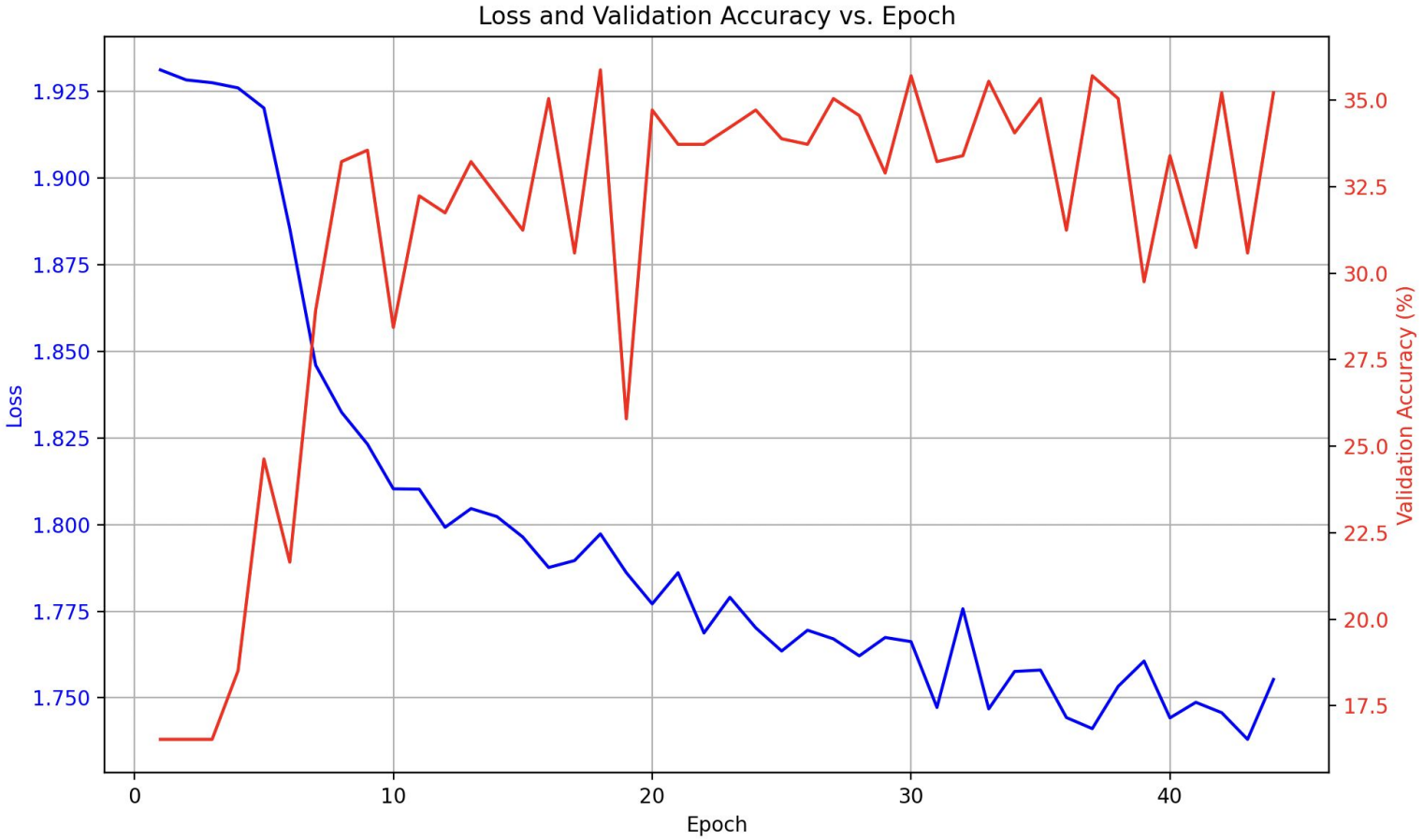
Dense

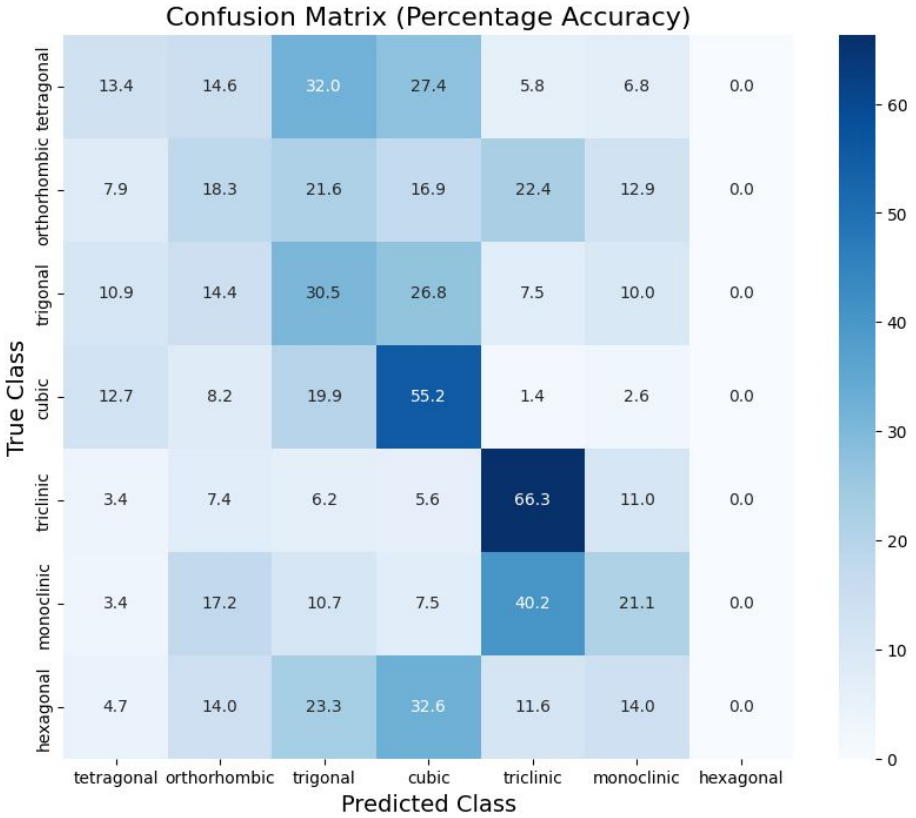

8500
input
4000
relu
dropout, 0.6
dense
3000
relu
dropout, 0.5
dense
1000
relu
dropout, 0.4
dense
800
relu
dropout, 0.3
dense
7
softmax
dense

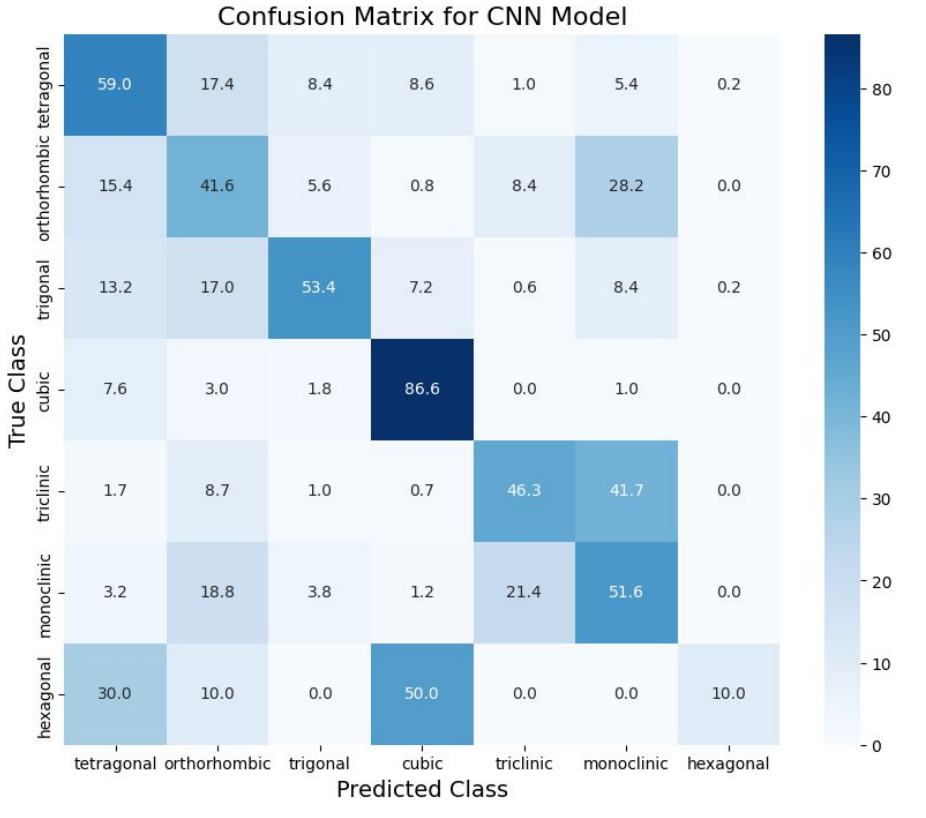
Dense Model Accuracy

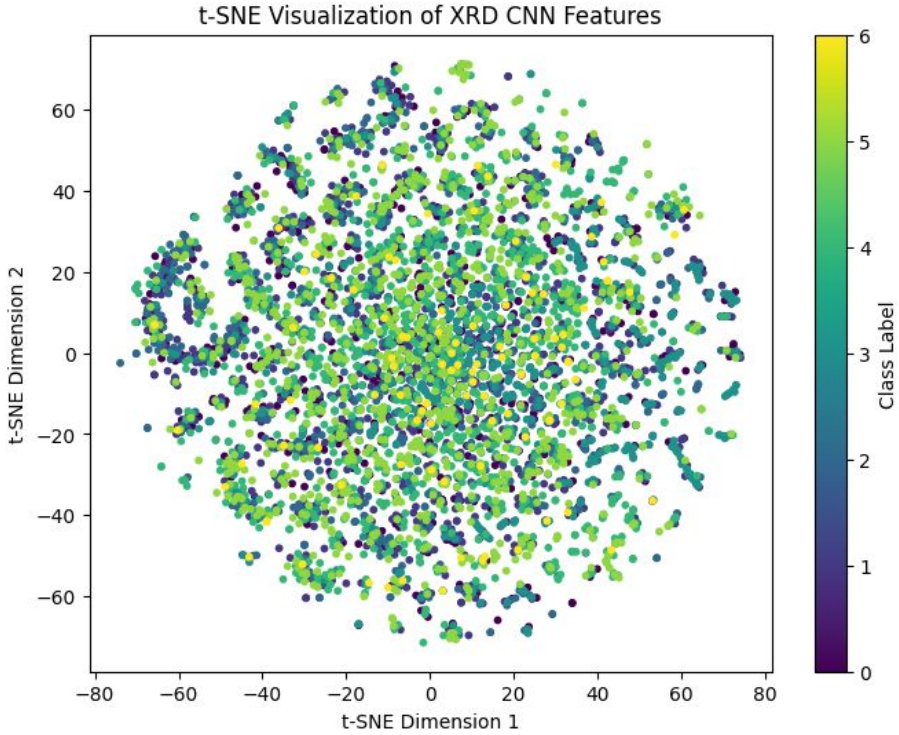
All data 35.25 % **Peak data** 56.4 %

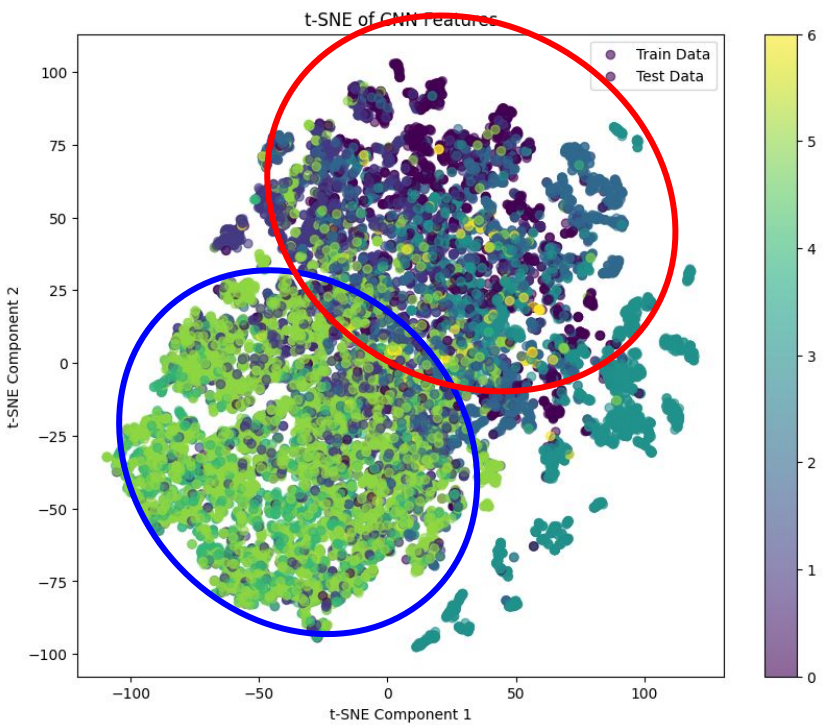
Der


Loss and Validation Accuracy vs. Epoch
Loss
Validation Accuracy (%)
Epoch
1.925
1.900
1.875
1.850
1.825
1.800
1.775
1.750
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
0
10
20
30
40

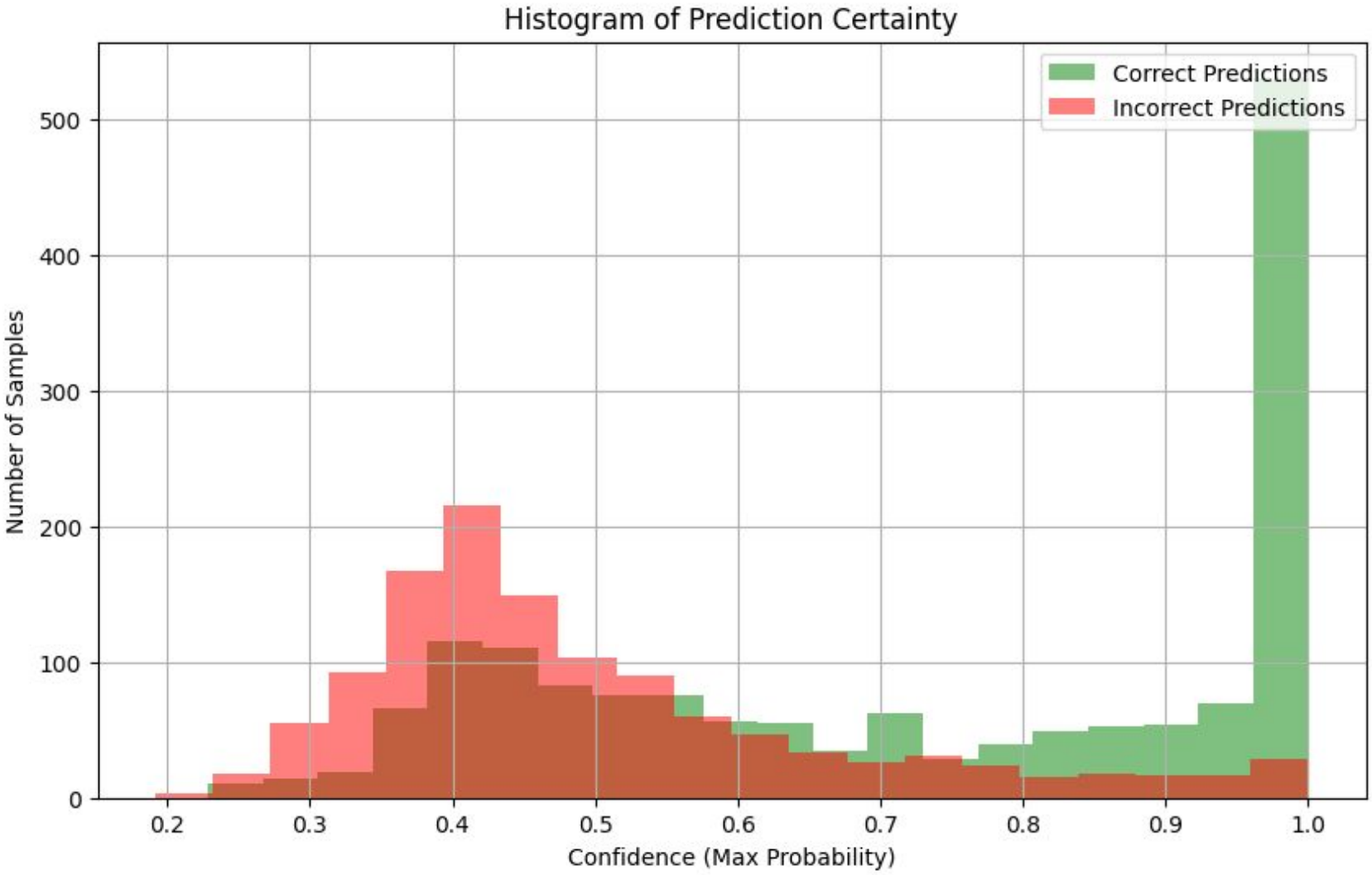

Confusion Matrix (Percentage Accuracy)
True Class
Predicted Class
tetragonal
orthorhombic
trigonal
cubic
triclinic
monoclinic
hexagonal
13.4 14.6 32.0 27.4 5.8 6.8 0.0
7.9 18.3 21.6 16.9 22.4 12.9 0.0
10.9 14.4 30.5 26.8 7.5 10.0 0.0
12.7 8.2 19.9 55.2 1.4 2.6 0.0
3.4 7.4 6.2 5.6 66.3 11.0 0.0
3.4 17.2 10.7 7.5 40.2 21.1 0.0
4.7 14.0 23.3 32.6 11.6 14.0 0.0
0
10
20
30
40
50
60


Confusion Matrix for CNN Model
True Class
Predicted Class
tetragonal
orthorhombic
trigonal
cubic
triclinic
monoclinic
hexagonal
59.0 17.4 8.4 8.6 1.0 5.4 0.2
15.4 41.6 5.6 0.8 8.4 28.2 0.0
13.2 17.0 53.4 7.2 0.6 8.4 0.2
7.6 3.0 1.8 86.6 0.0 1.0 0.0
1.7 8.7 1.0 0.7 46.3 41.7 0.0
3.2 18.8 3.8 1.2 21.4 51.6 0.0
30.0 10.0 0.0 50.0 0.0 0.0 10.0
0
10
20
30
40
50
60
70
80


t-SNE Visualization of XRD CNN Features
t-SNE Dimension 2
t-SNE Dimension 1
Class Label

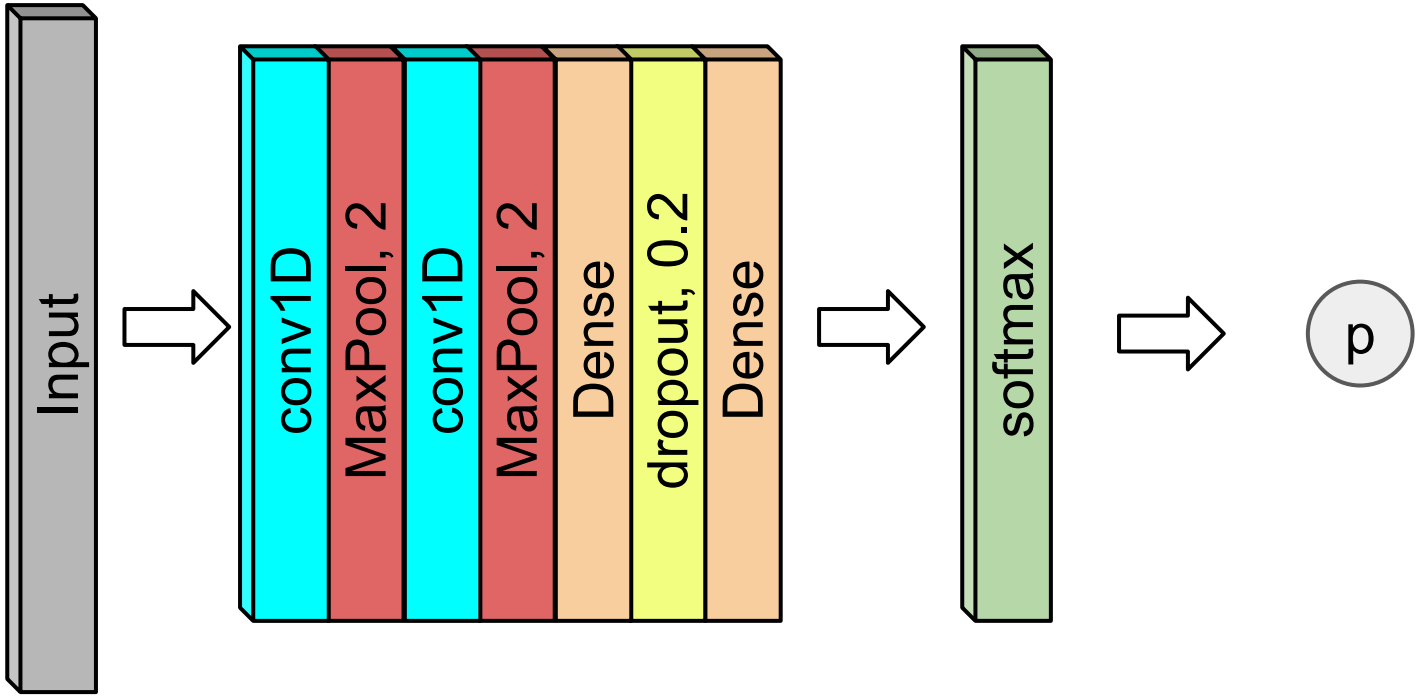

Train Data
Test Data
t-SNE Component 2
t-SNE Component 1

D


Histogram of Prediction Certainty
Correct Predictions
Incorrect Predictions
Number of Samples
0
100
200
300
400
500
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Confidence (Max Probability)

2. CNN + Dense

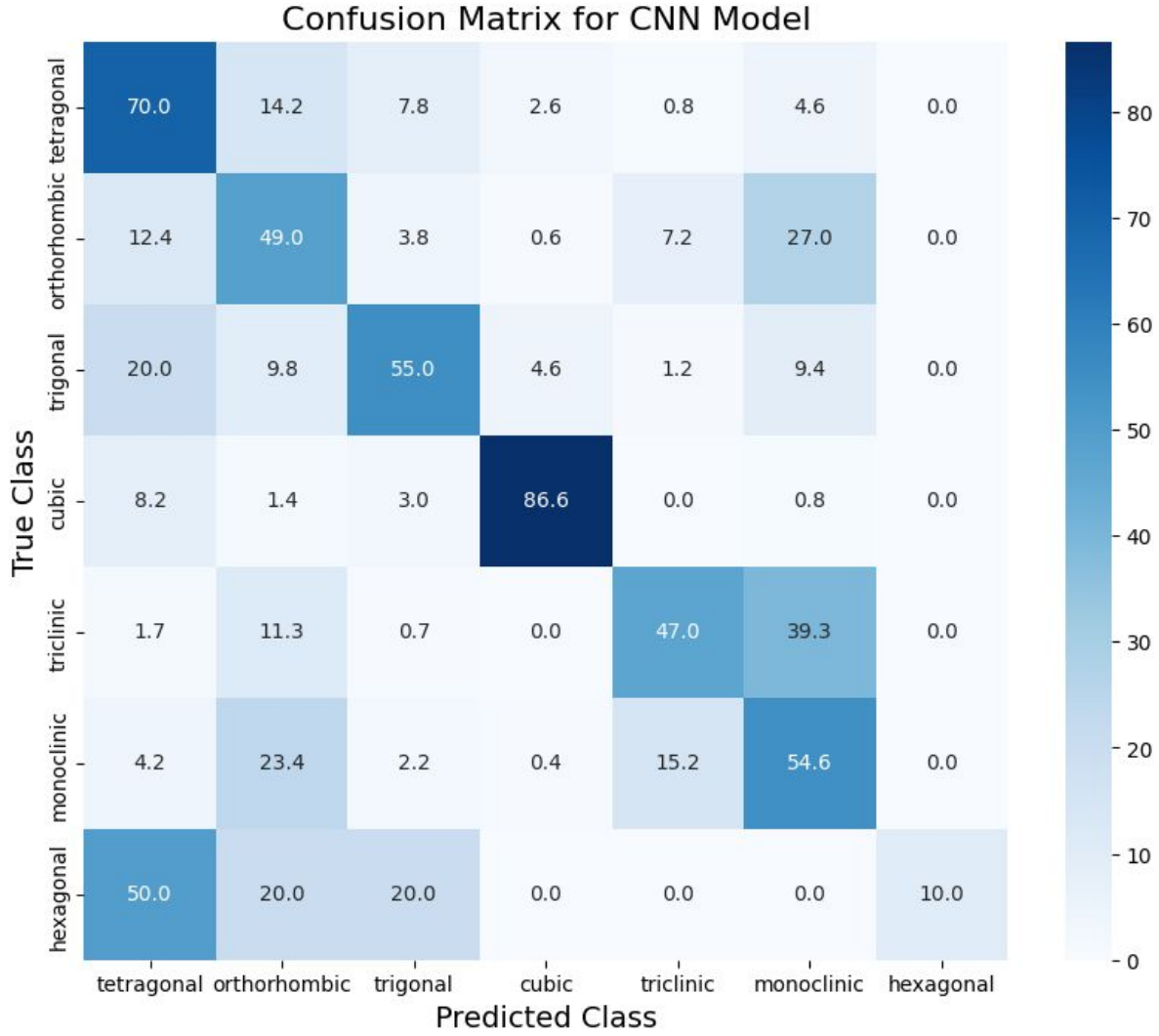
CNN-Dense: Architecture



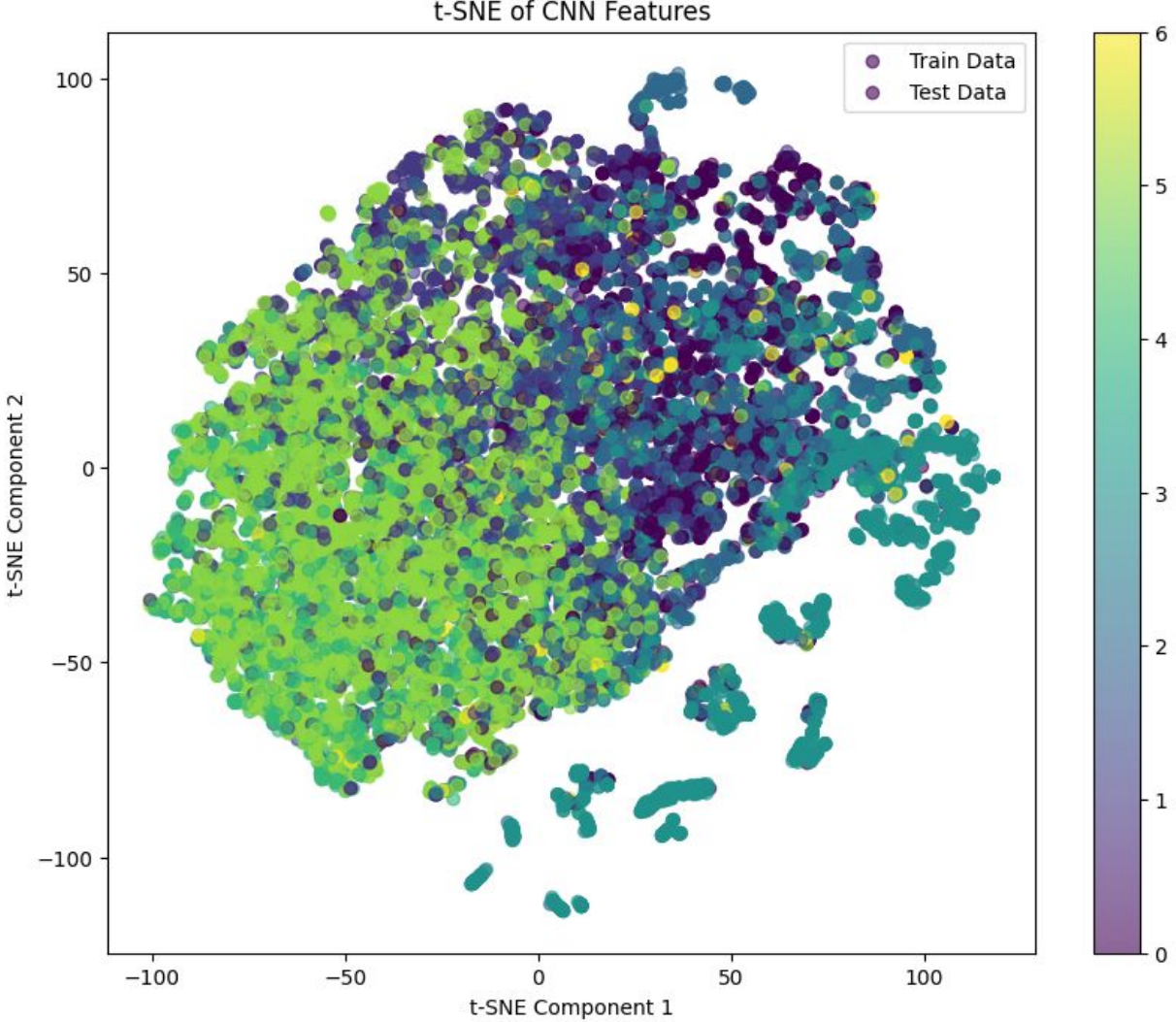
CNN + Dense Model Accuracy

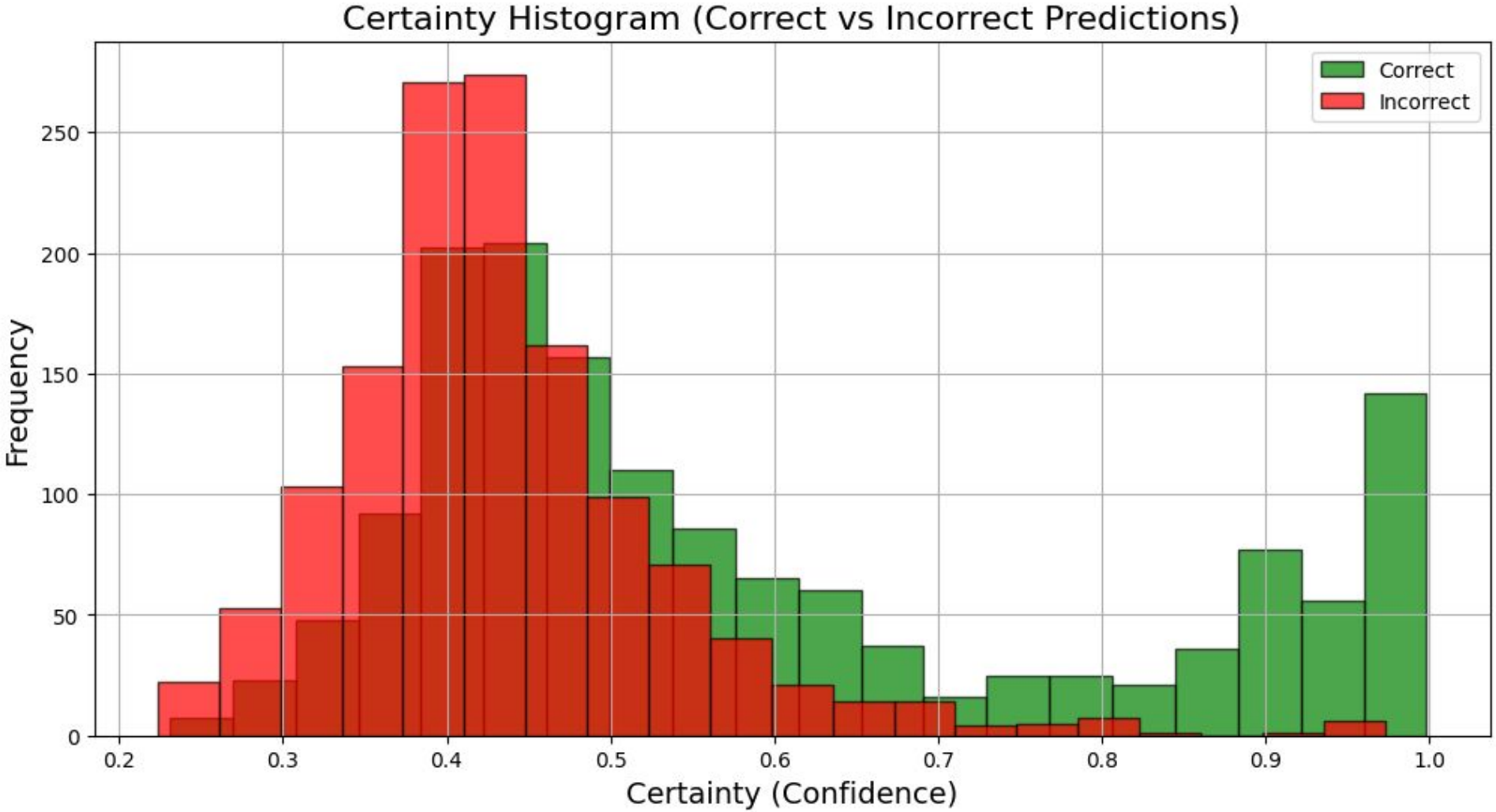
Peak data 60.37 %

CNN+De

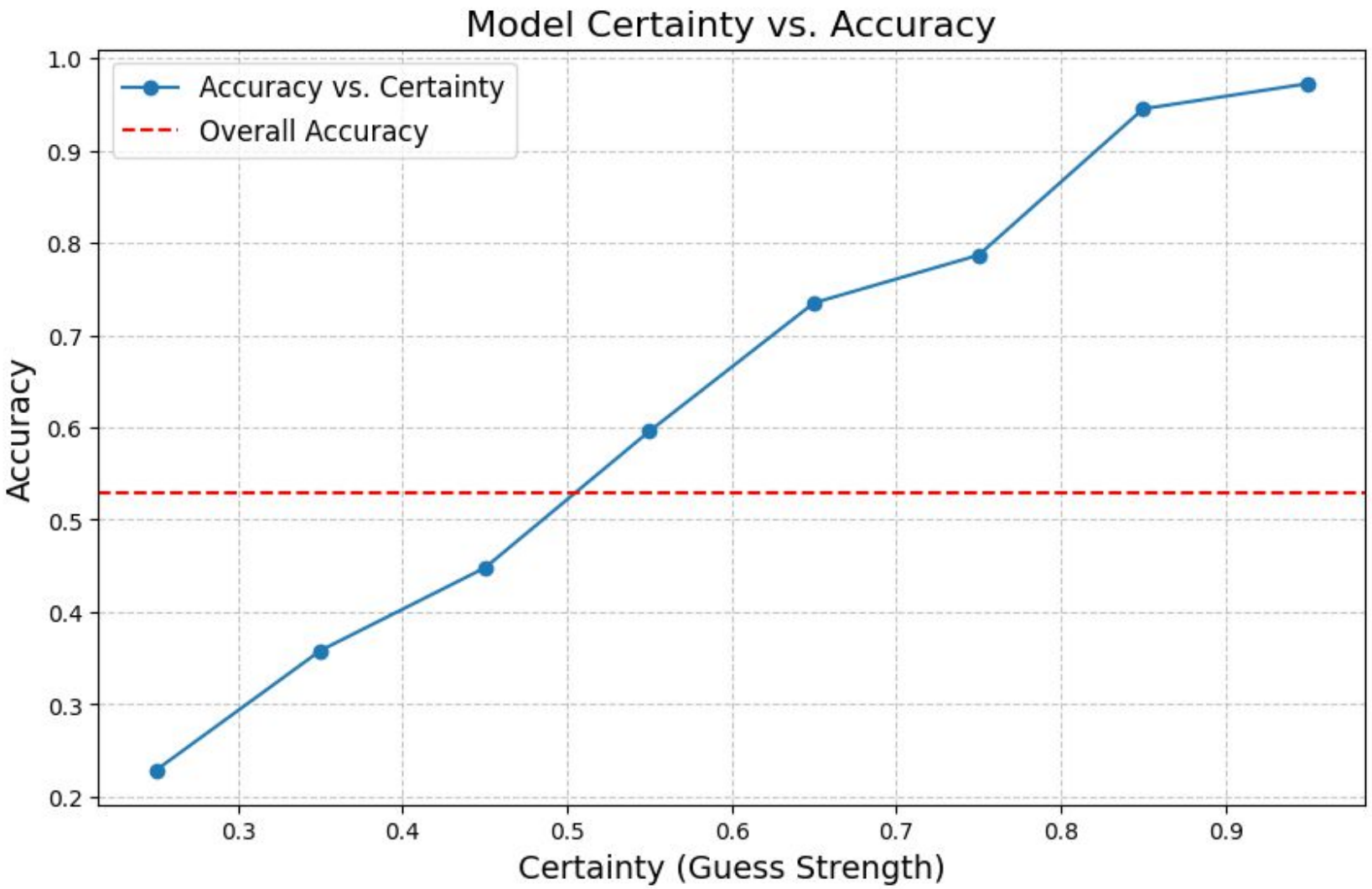

Confusion Matrix for CNN Model
True Class
Predicted Class
tetragonal
orthorhombic
trigonal
cubic
triclinic
monoclinic
hexagonal
70.0 14.2 7.8 2.6 0.8 4.6 0.0
12.4 49.0 3.8 0.6 7.2 27.0 0.0
20.0 9.8 55.0 4.6 1.2 9.4 0.0
8.2 1.4 3.0 86.6 0.0 0.8 0.0
1.7 11.3 0.7 0.0 47.0 39.3 0.0
4.2 23.4 2.2 0.4 15.2 54.6 0.0
50.0 20.0 20.0 0.0 0.0 0.0 10.0
80
70
60
50
40
30
20
10
0

CNN+De

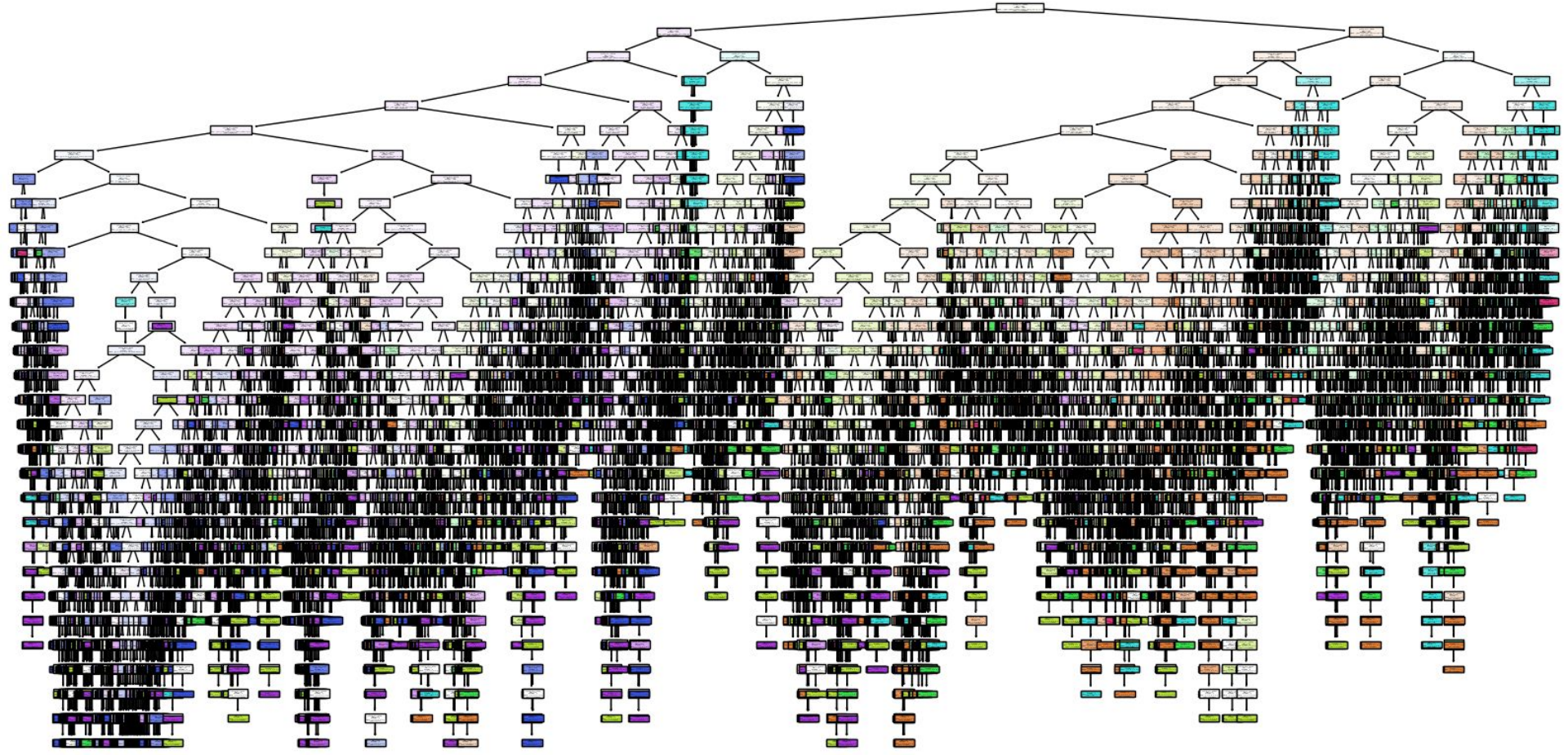

t-SNE of CNN Features
Train Data
Test Data
t-SNE Component 2
t-SNE Component 1
100
50
0
−50
−100
−100
−50
0
50
100
6
5
4
3
2
1
0


Certainty Histogram (Correct vs Incorrect Predictions)
Correct
Incorrect
Frequency
0
50
100
150
200
250
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Certainty (Confidence)

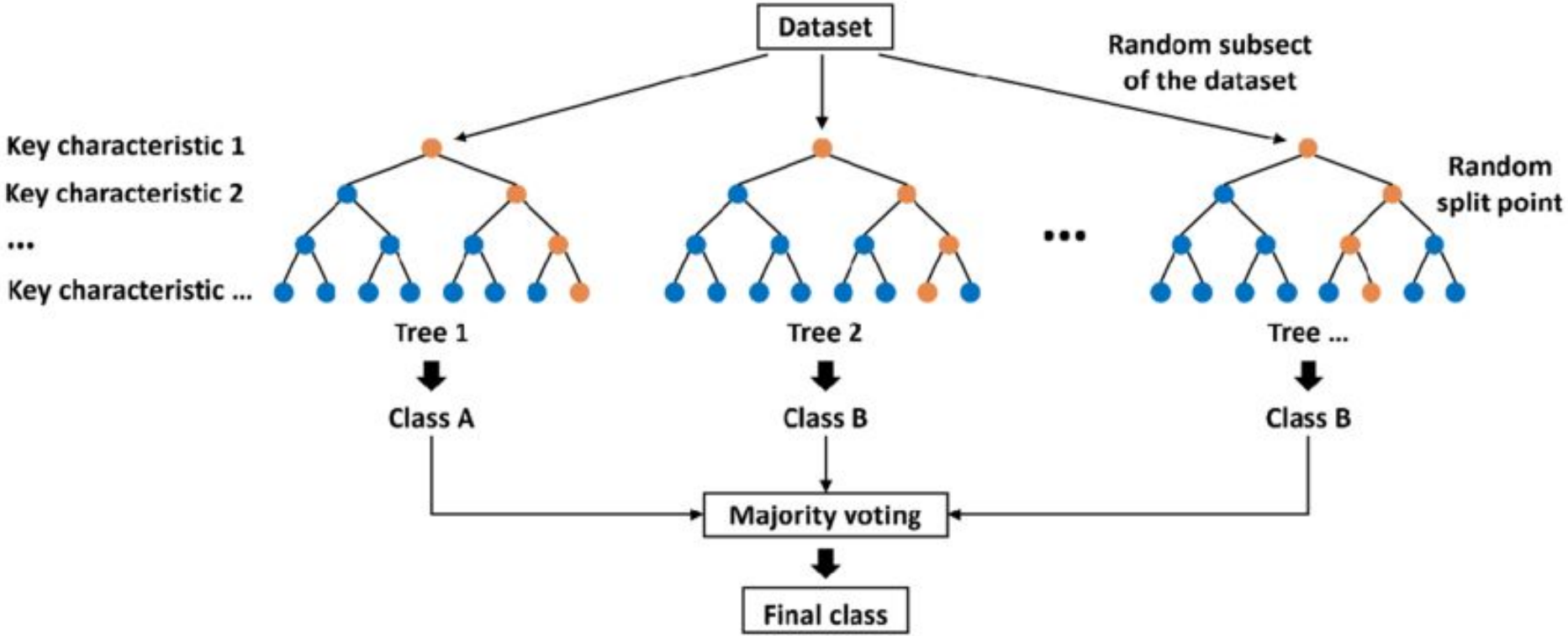
Ex


Model Certainty vs. Accuracy
Accuracy vs. Certainty
Overall Accuracy
Accuracy
Certainty (Guess Strength)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

3. Extreme Randomized Trees



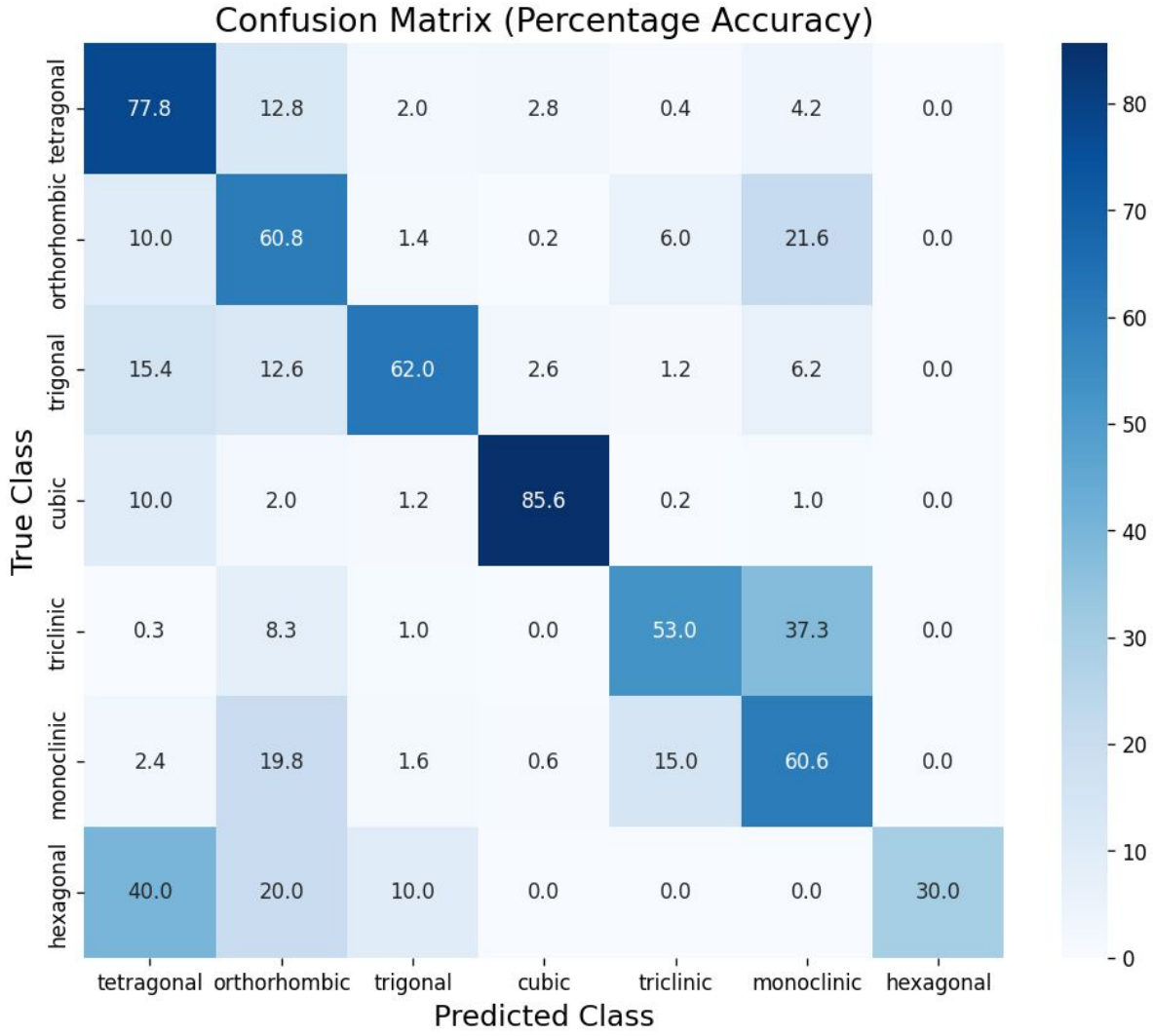
Extra Trees Classifier - Peak



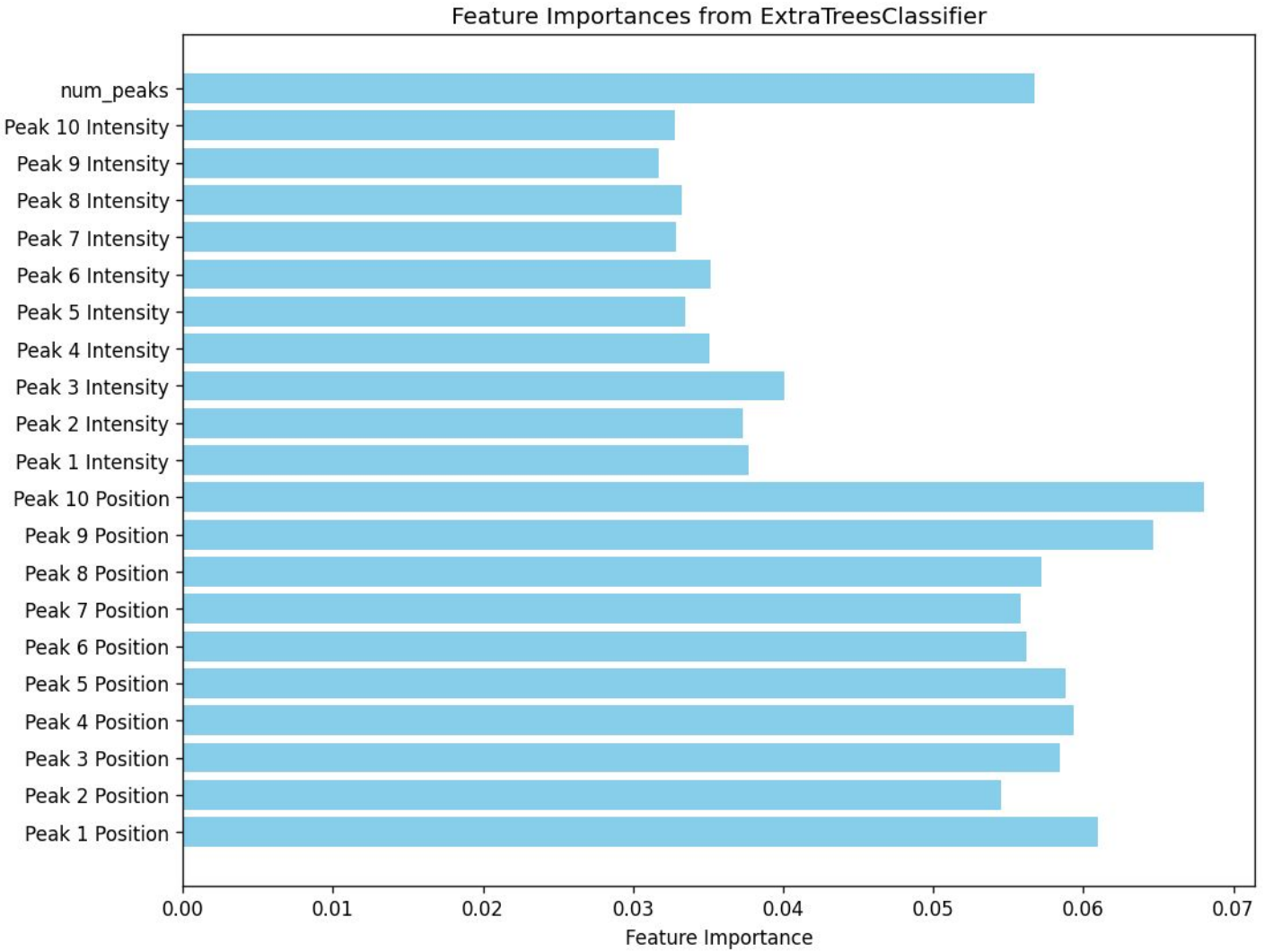
Extra-Trees Accuracy

Peak data 67.73 %

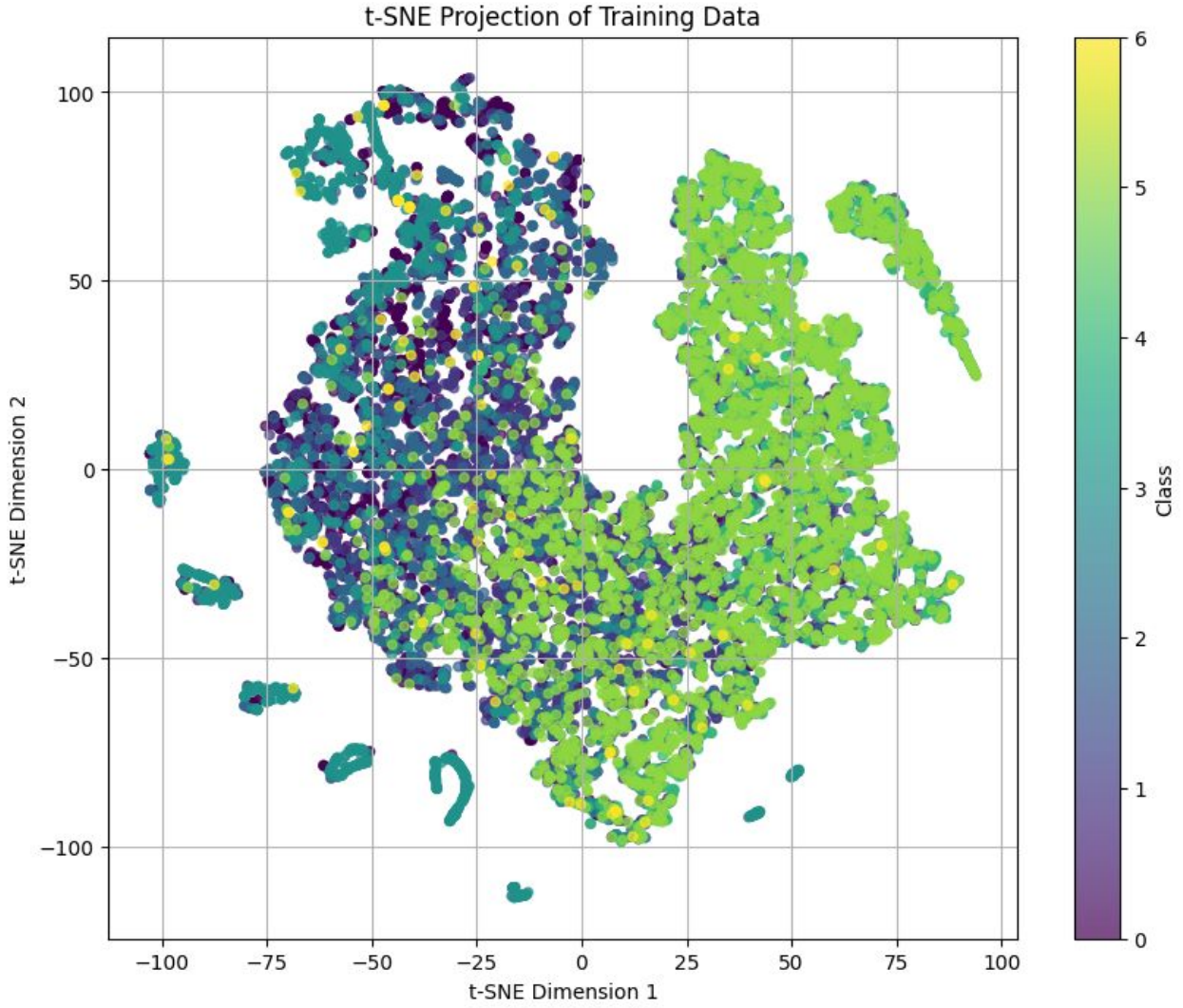
Extra Tre

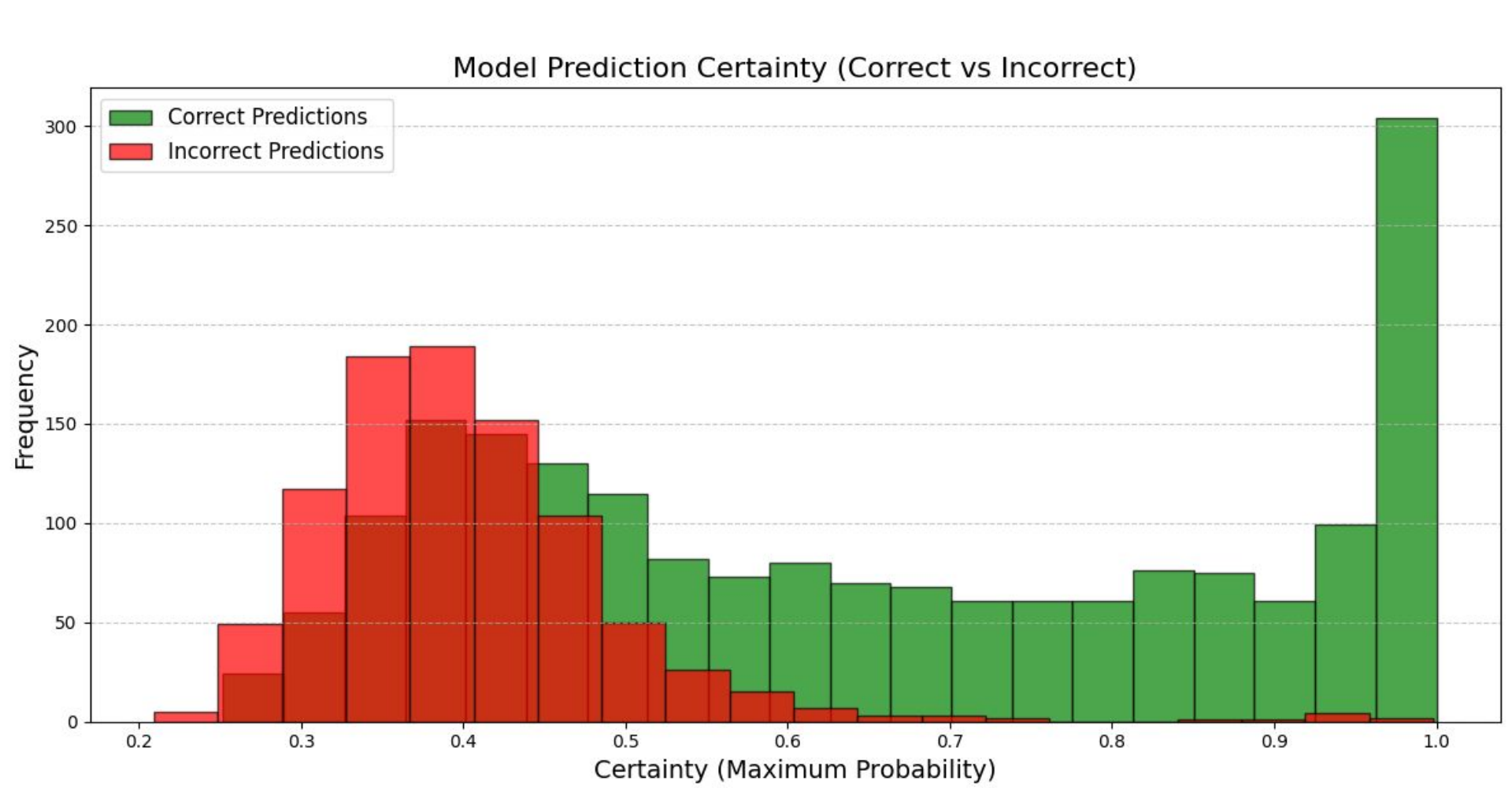

Confusion Matrix (Percentage Accuracy)
True Class
Predicted Class
tetragonal
orthorhombic
trigonal
cubic
triclinic
monoclinic
hexagonal
77.8 12.8 2.0 2.8 0.4 4.2 0.0
10.0 60.8 1.4 0.2 6.0 21.6 0.0
15.4 12.6 62.0 2.6 1.2 6.2 0.0
10.0 2.0 1.2 85.6 0.2 1.0 0.0
0.3 8.3 1.0 0.0 53.0 37.3 0.0
2.4 19.8 1.6 0.6 15.0 60.6 0.0
40.0 20.0 10.0 0.0 0.0 0.0 30.0
80
70
60
50
40
30
20
10
0

Extra

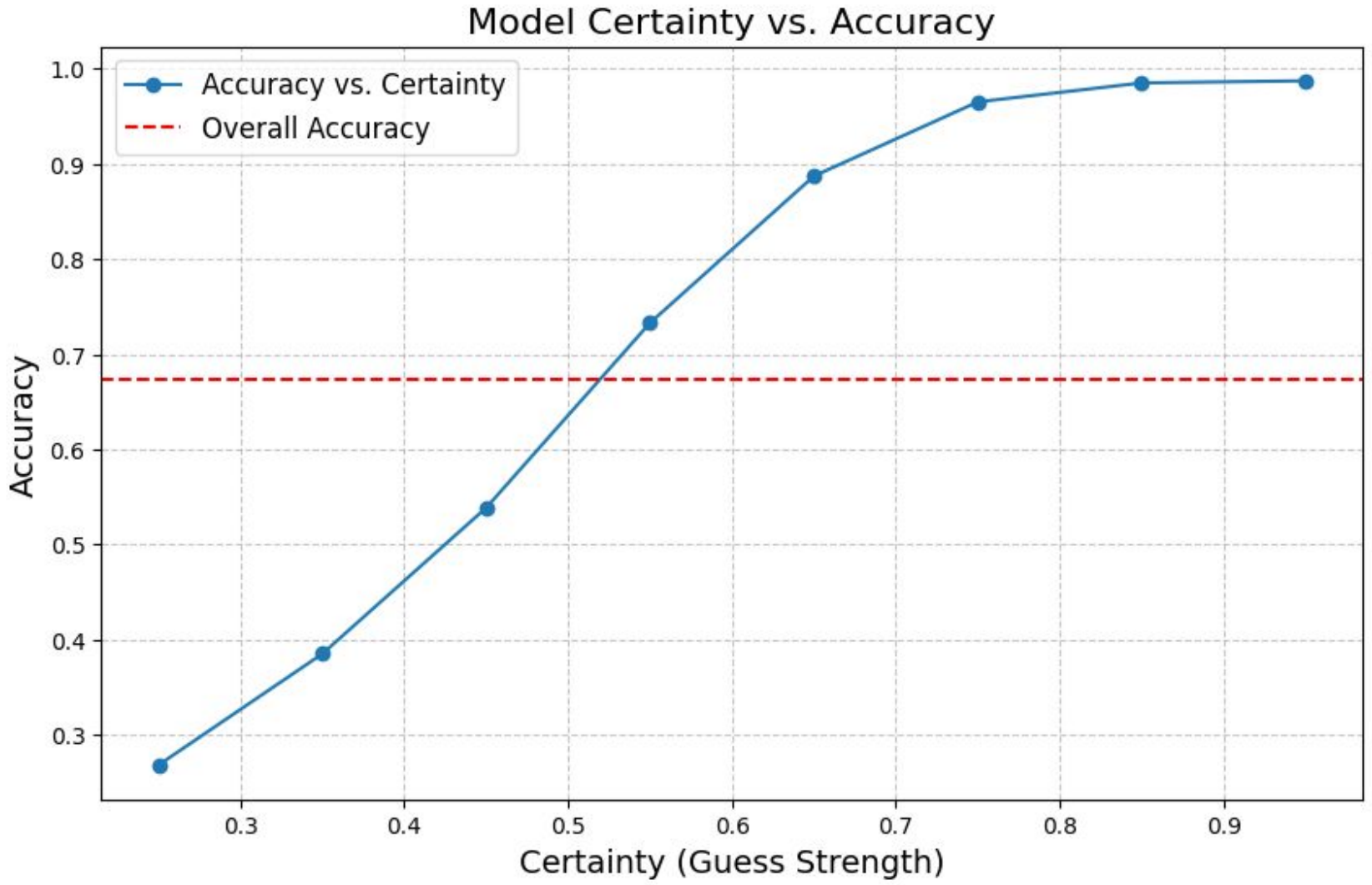

Feature Importances from ExtraTreesClassifier
num_peaks
Peak 10 Intensity
Peak 9 Intensity
Peak 8 Intensity
Peak 7 Intensity
Peak 6 Intensity
Peak 5 Intensity
Peak 4 Intensity
Peak 3 Intensity
Peak 2 Intensity
Peak 1 Intensity
Peak 10 Position
Peak 9 Position
Peak 8 Position
Peak 7 Position
Peak 6 Position
Peak 5 Position
Peak 4 Position
Peak 3 Position
Peak 2 Position
Peak 1 Position
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Feature Importance

Extra Tre


t-SNE Projection of Training Data
t-SNE Dimension 2
t-SNE Dimension 1
Class
100
50
0
−50
−100
−100
−75
−50
−25
0
25
50
75
100
6
5
4
3
2
1
0

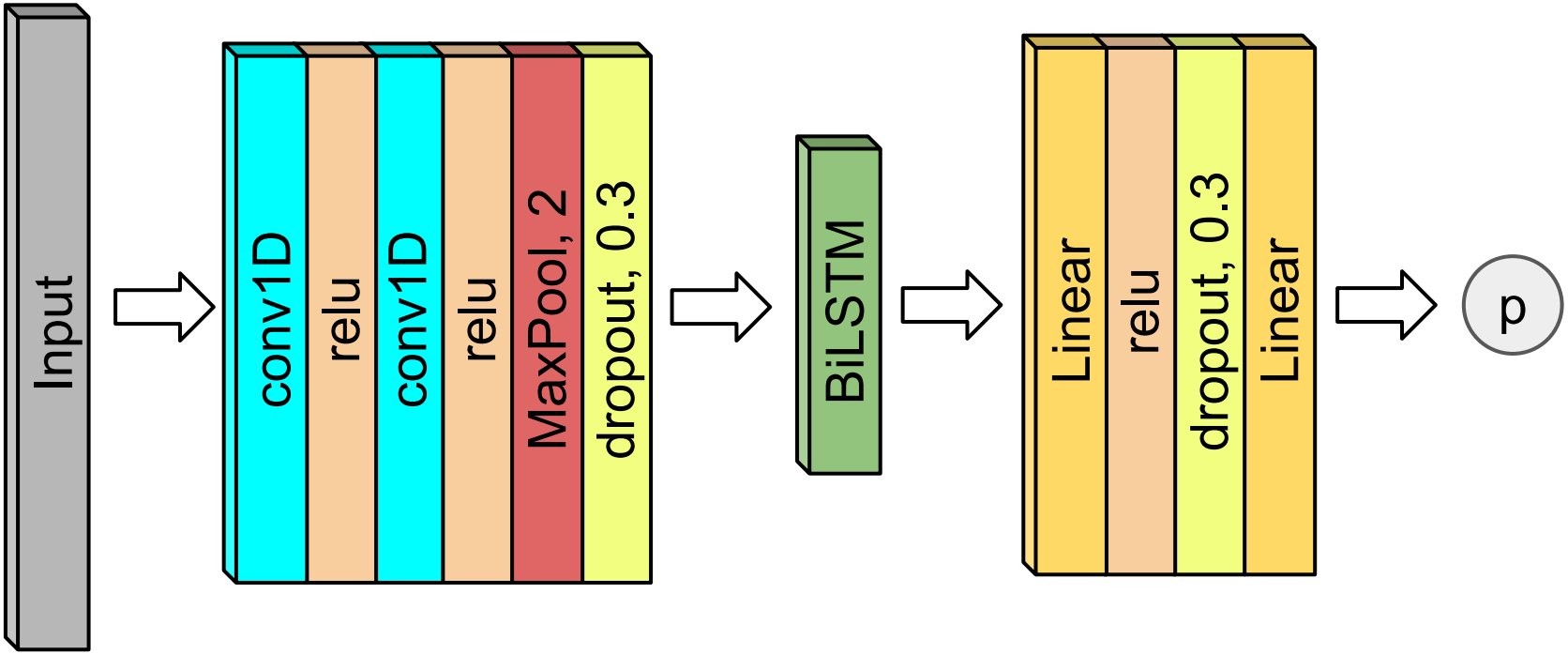

Model Prediction Certainty (Correct vs Incorrect)
Correct Predictions
Incorrect Predictions
Frequency
0
50
100
150
200
250
300
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Certainty (Maximum Probability)

E

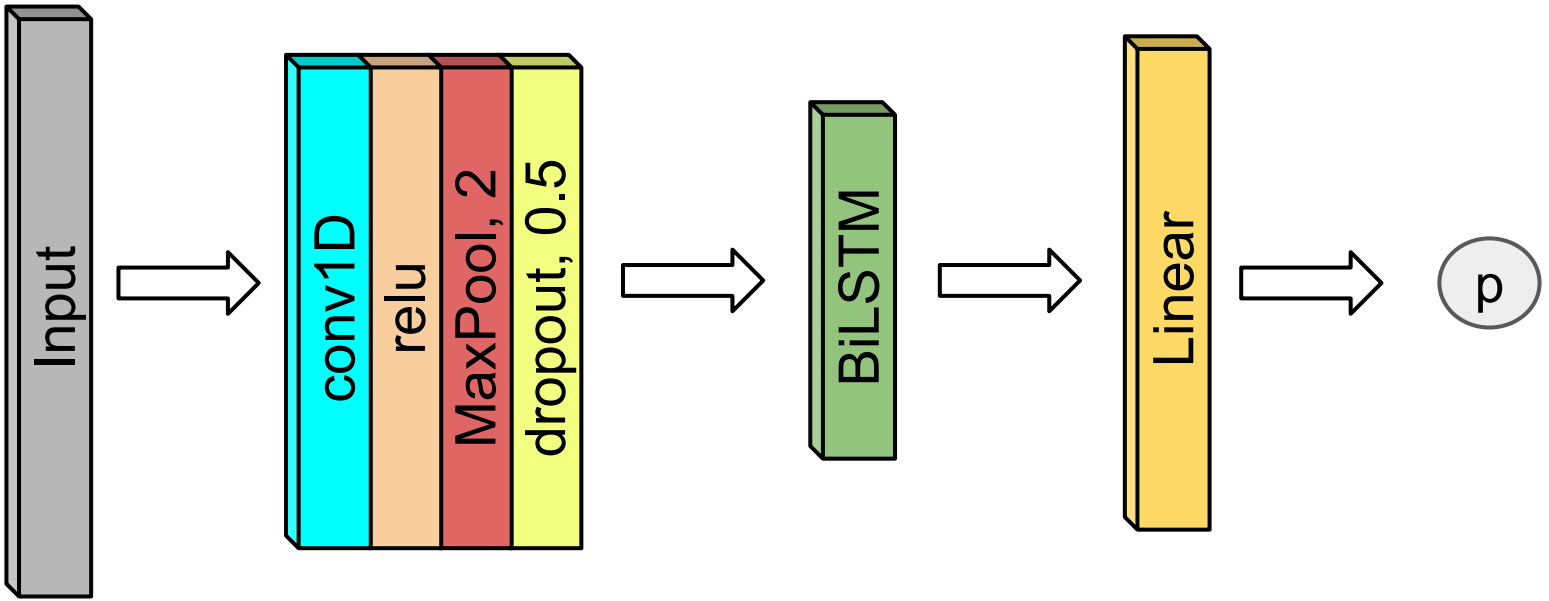

Model Certainty vs. Accuracy
Accuracy vs. Certainty
Overall Accuracy
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Certainty (Guess Strength)

4. CNN + LSTM

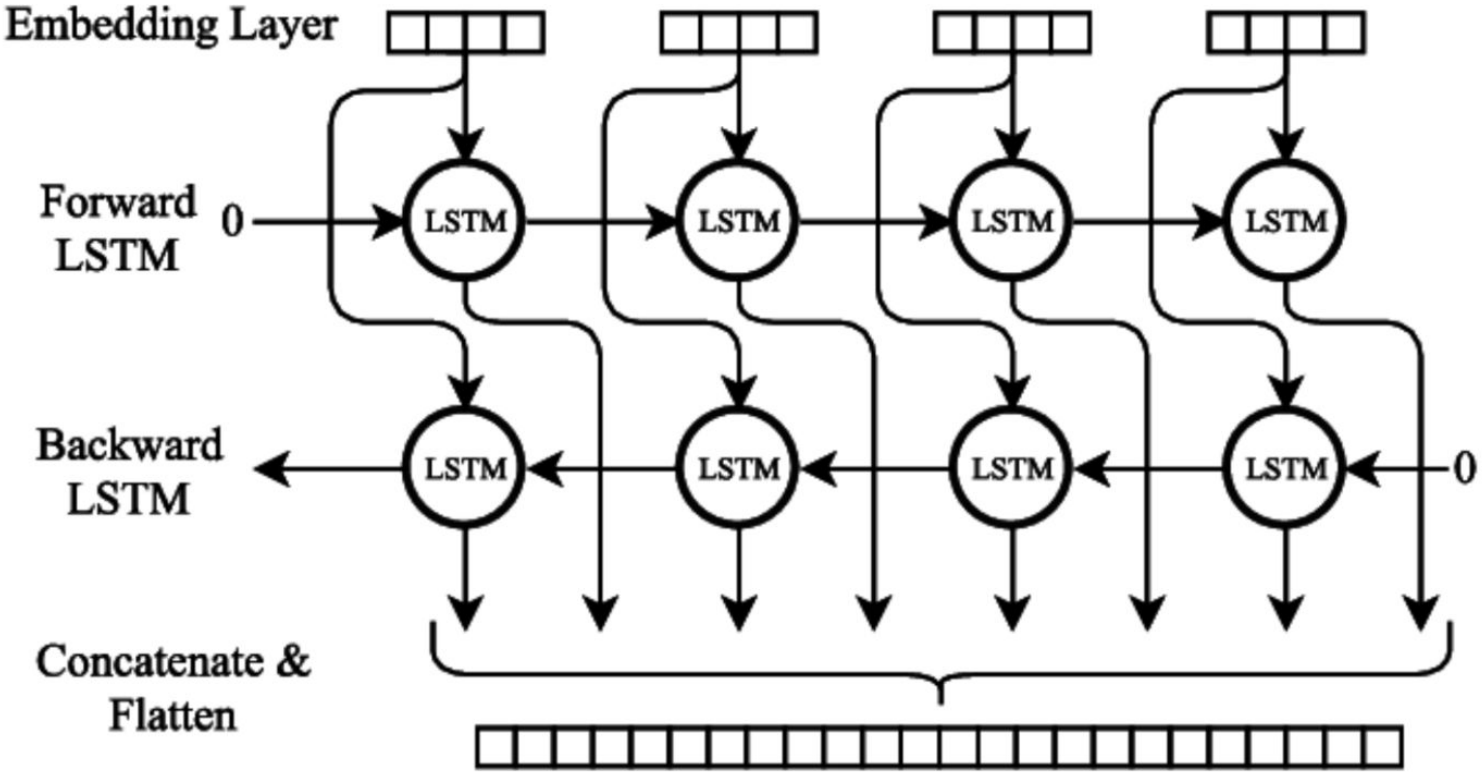
CNN-LSTM: Architecture



sCNN-LSTM: Architecture



BiLSTM: Architecture



Focal Loss

$$CE = -\log(p_t)$$

$$FL = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)$$

Focal Loss

$$FL = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)$$

weighting coefficient, set = 2

Models' Accuracy

All data:

sCNN-LSTM, CE/ FL: 52.66%, 50.47%

Peak data:

CNN-LSTM, CE/ FL: 64.65%, 54.88%

sCNN-LSTM, CE/ FL: 43.19%, 42.96%

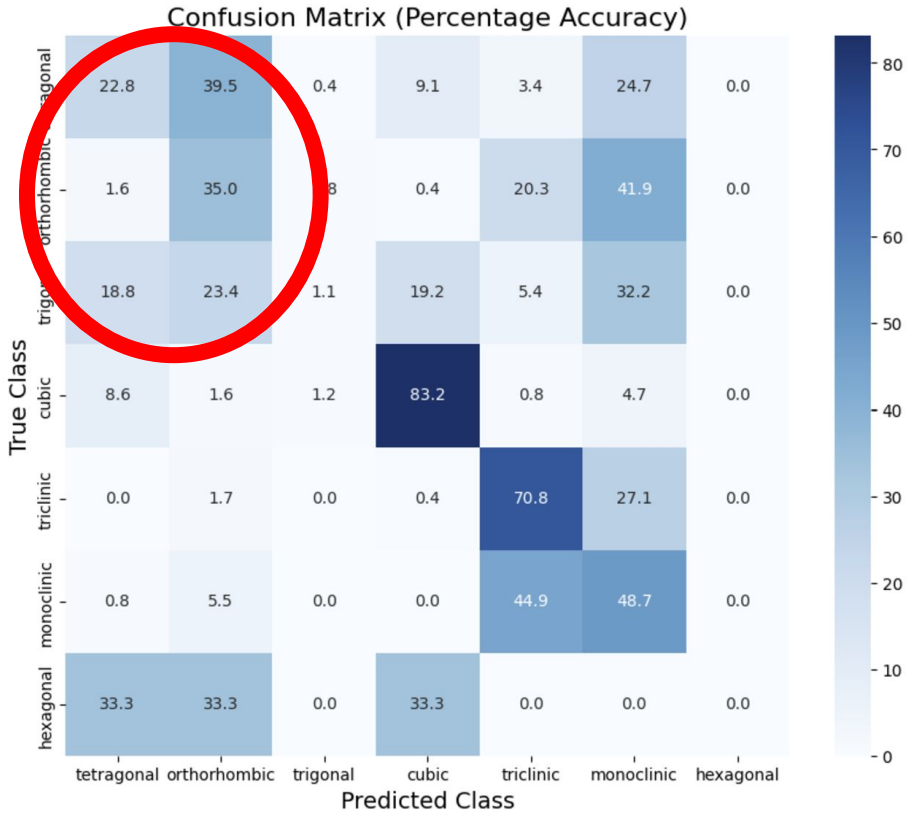
6 models

Comparison:

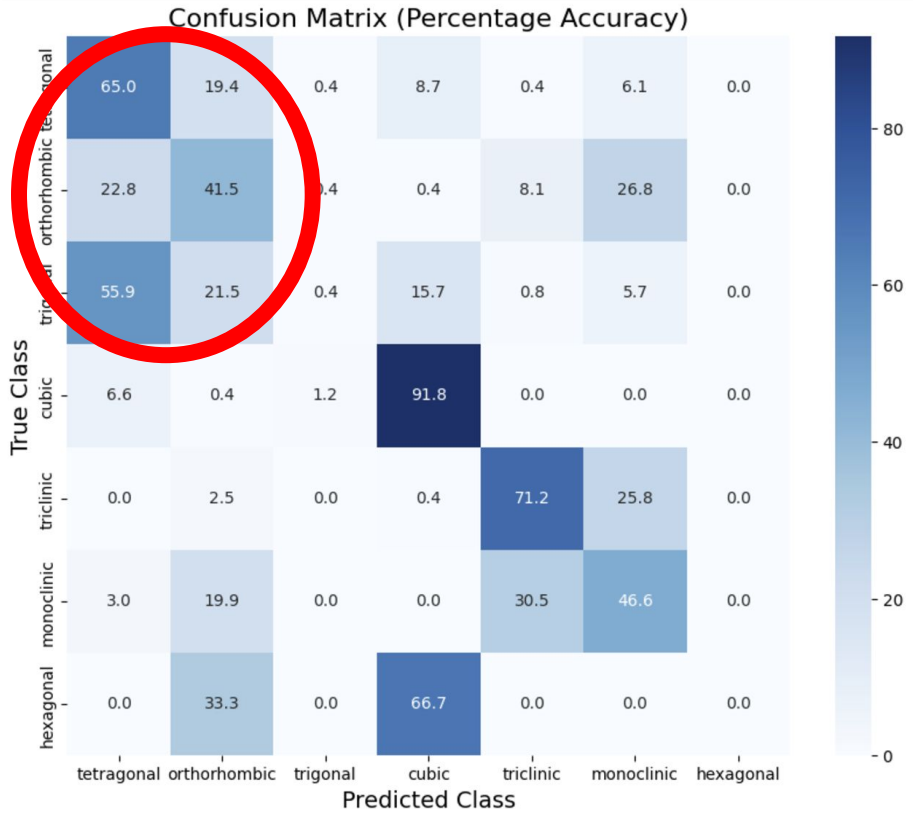
Peak Data 43.19% vs. All Data 52.66%

sCNN-LSTM, Cross Entropy

Peak Data/ All Data - Confusion Matrix

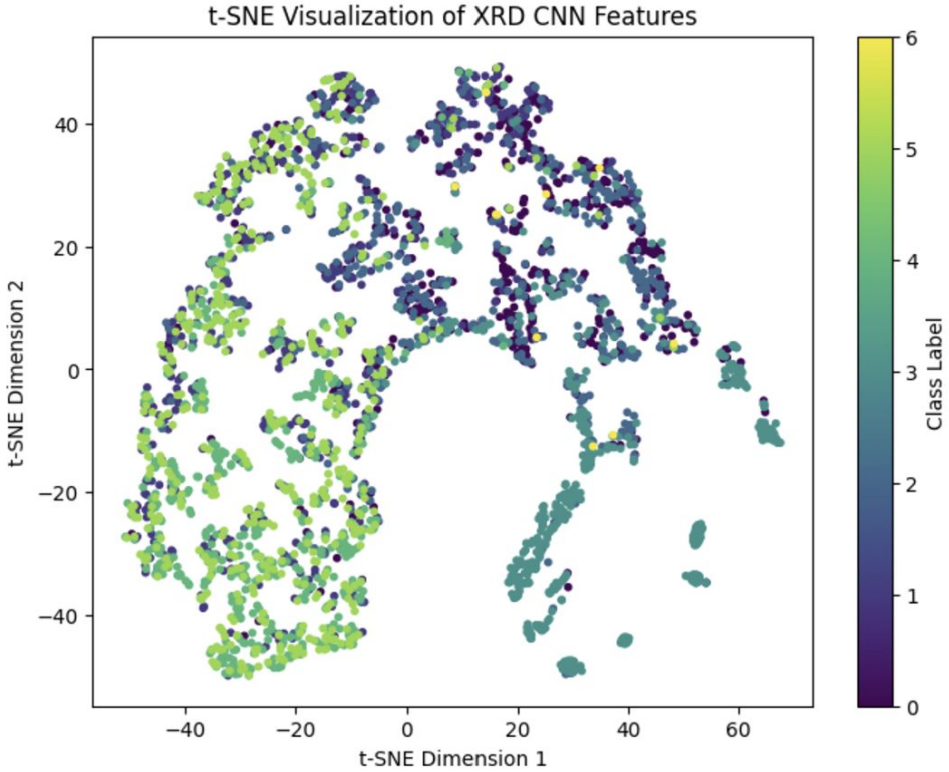


Peak Data

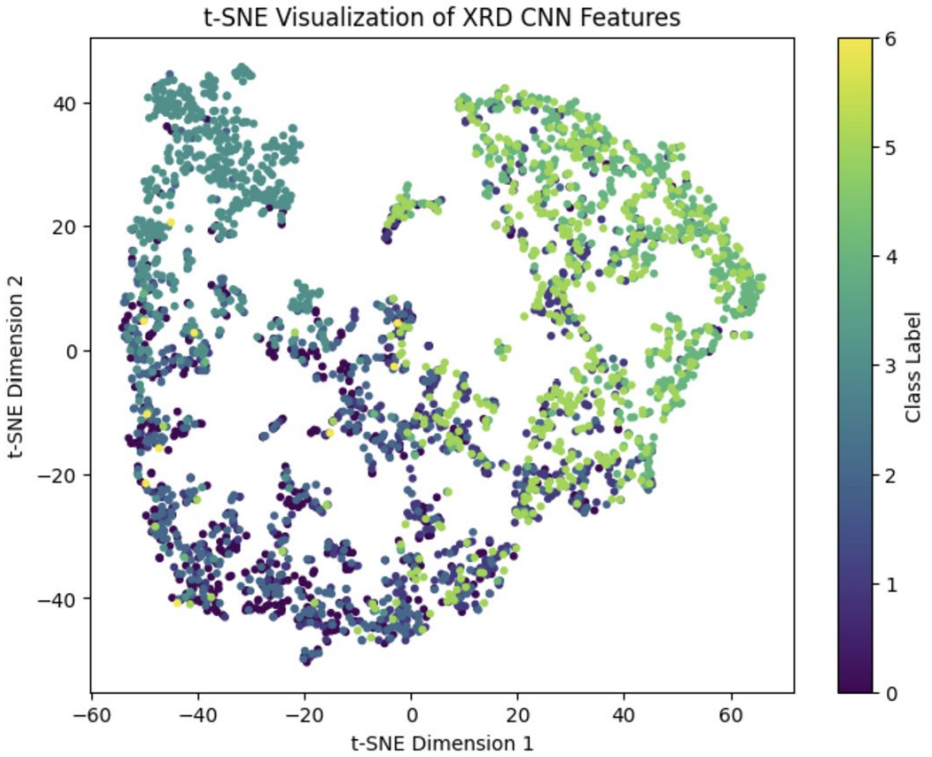


All Data

Peak Data/ All Data - TSNE

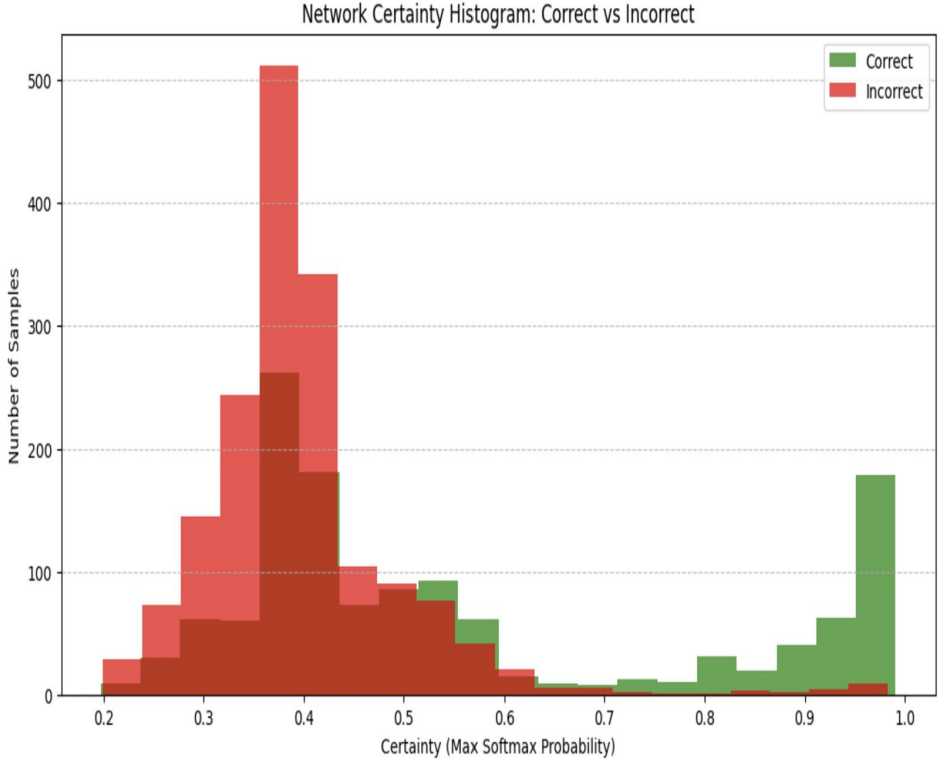


Peak Data

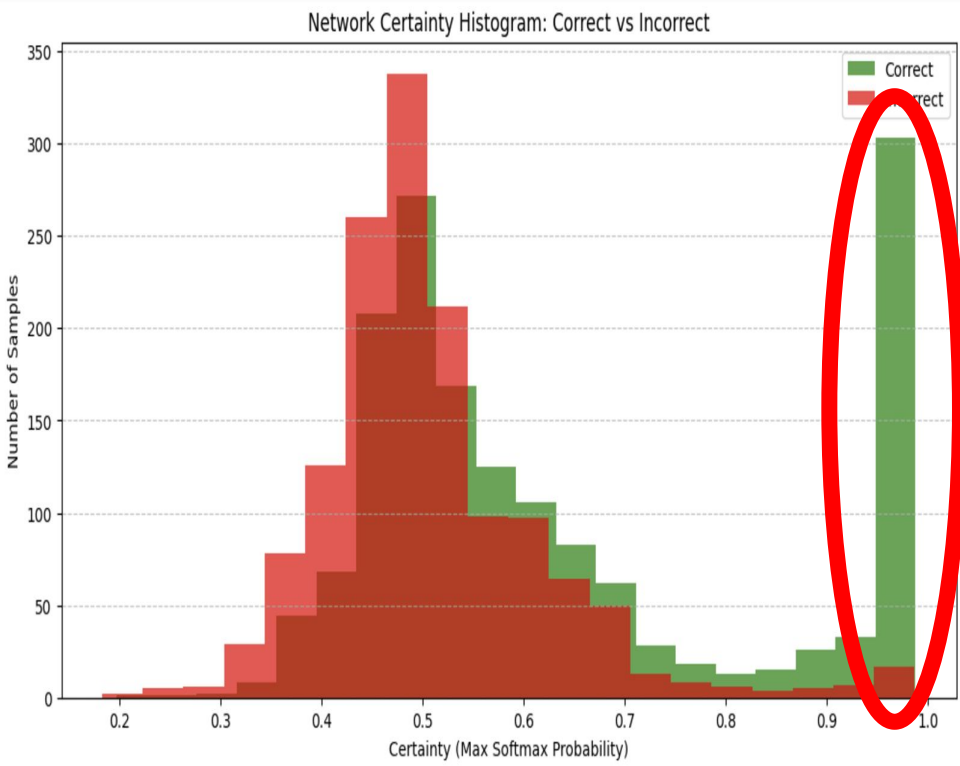


All Data

Peak Data/ All Data - Network Certainty



Peak Data



All Data

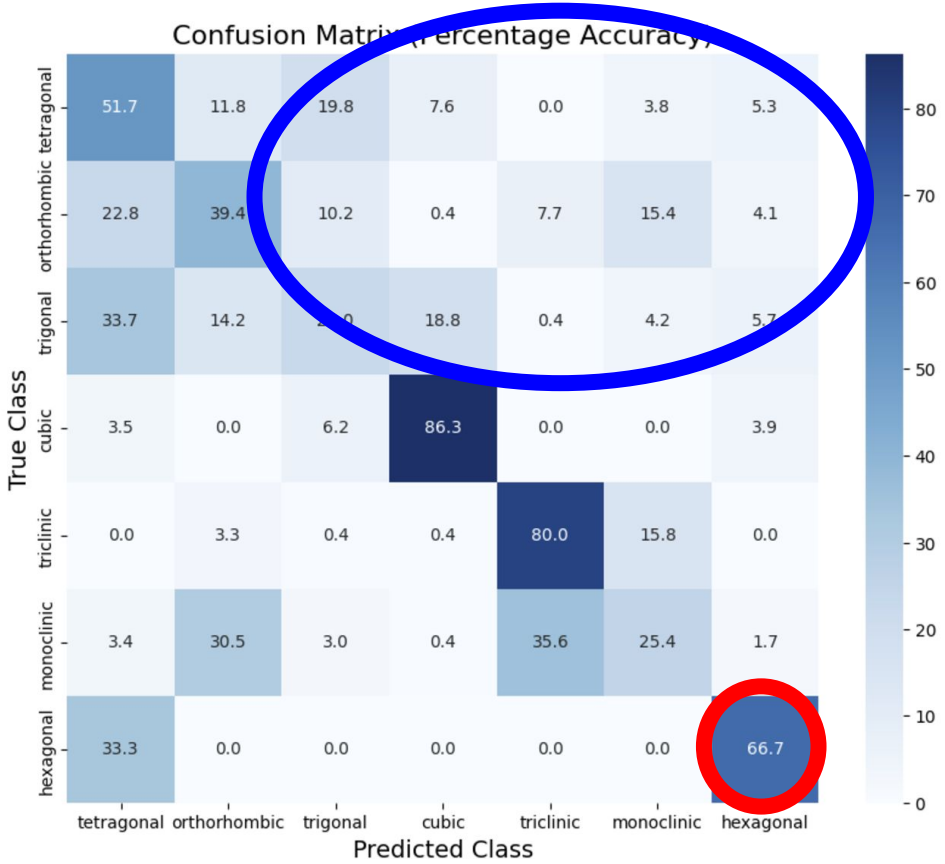
Comparison:

Cross Entropy 52.66% vs. Focal Loss 50.47%

sCNN-LSTM, All Data

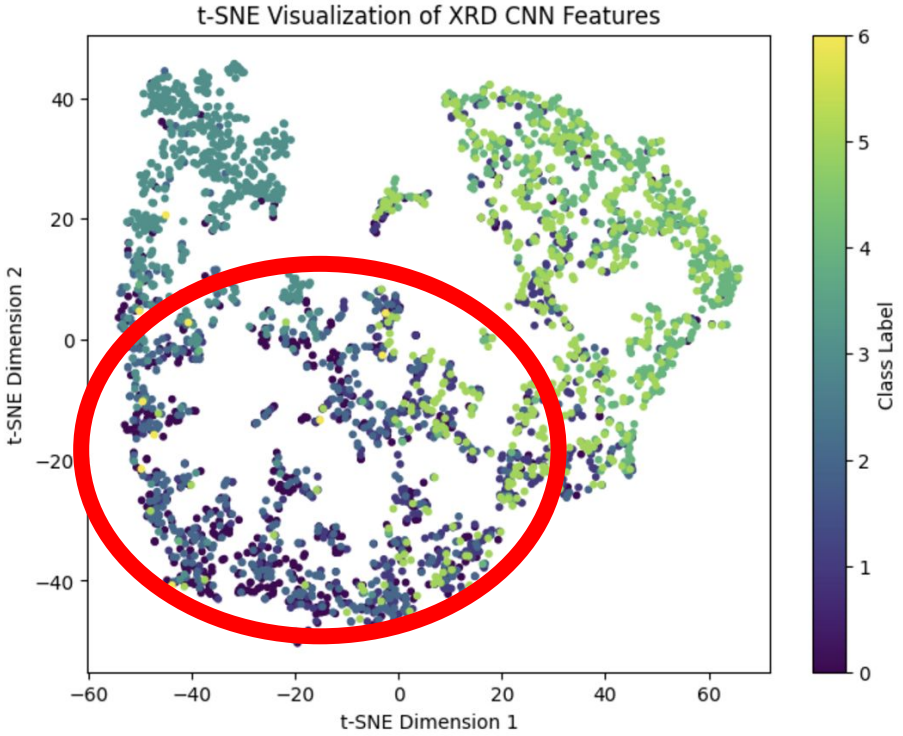
CE/ FL - Confusion Matrix

Cross Entropy

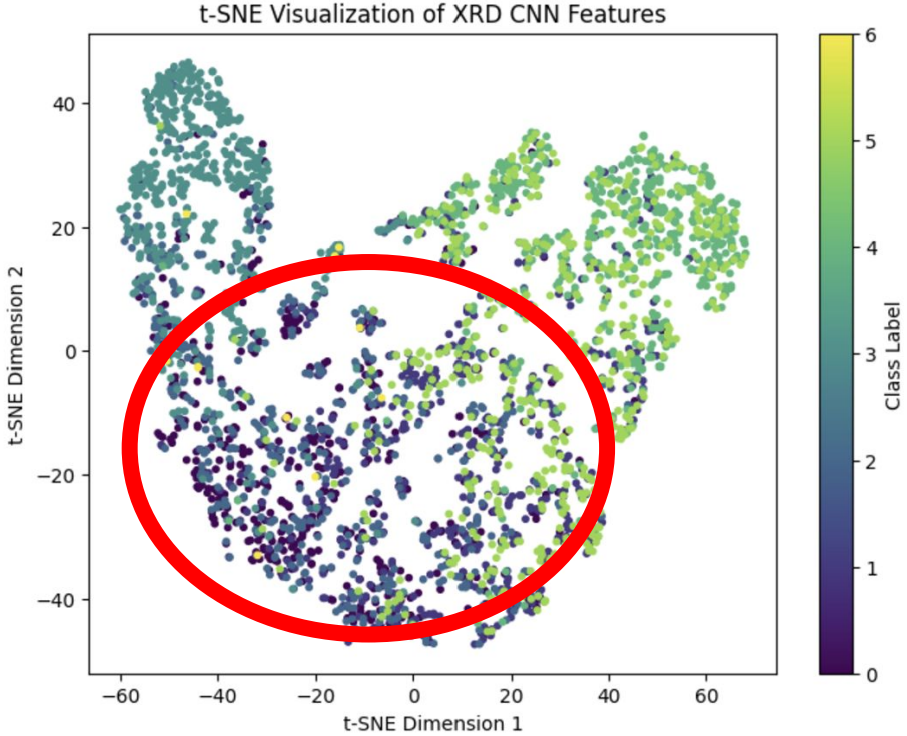


Focal Loss

CE/ FL - TSNE

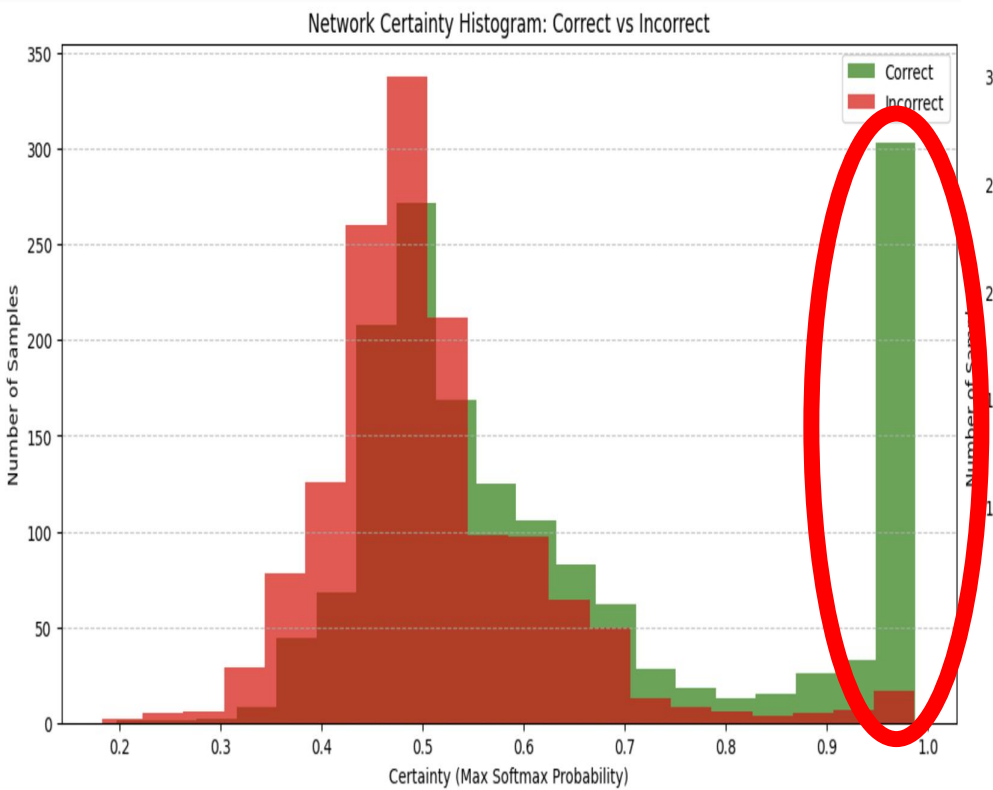


Cross Entropy

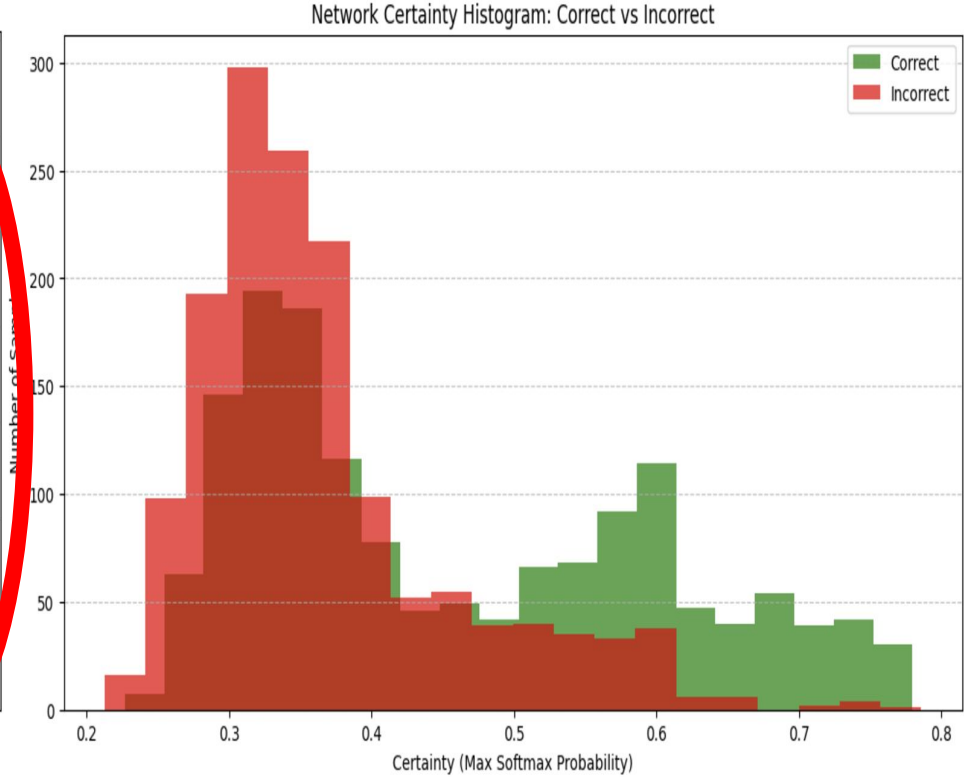


Focal Loss

CE/ FL - Network Certainty



Cross Entropy



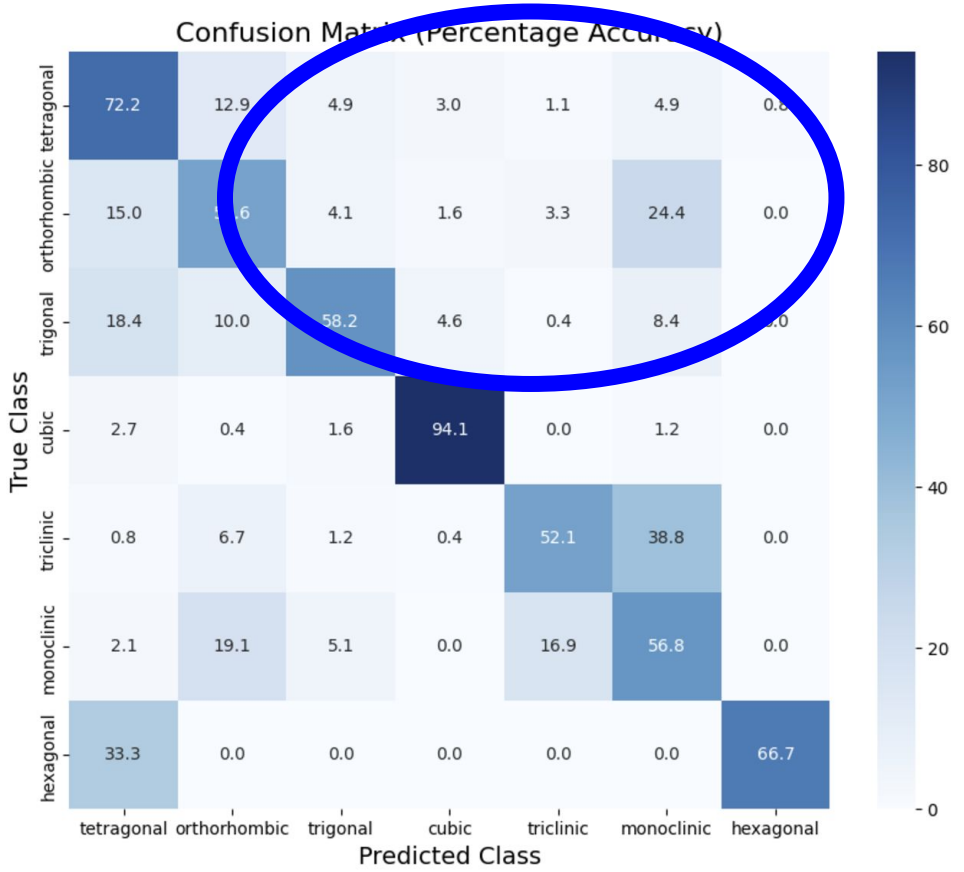
Focal Loss

Comparison:

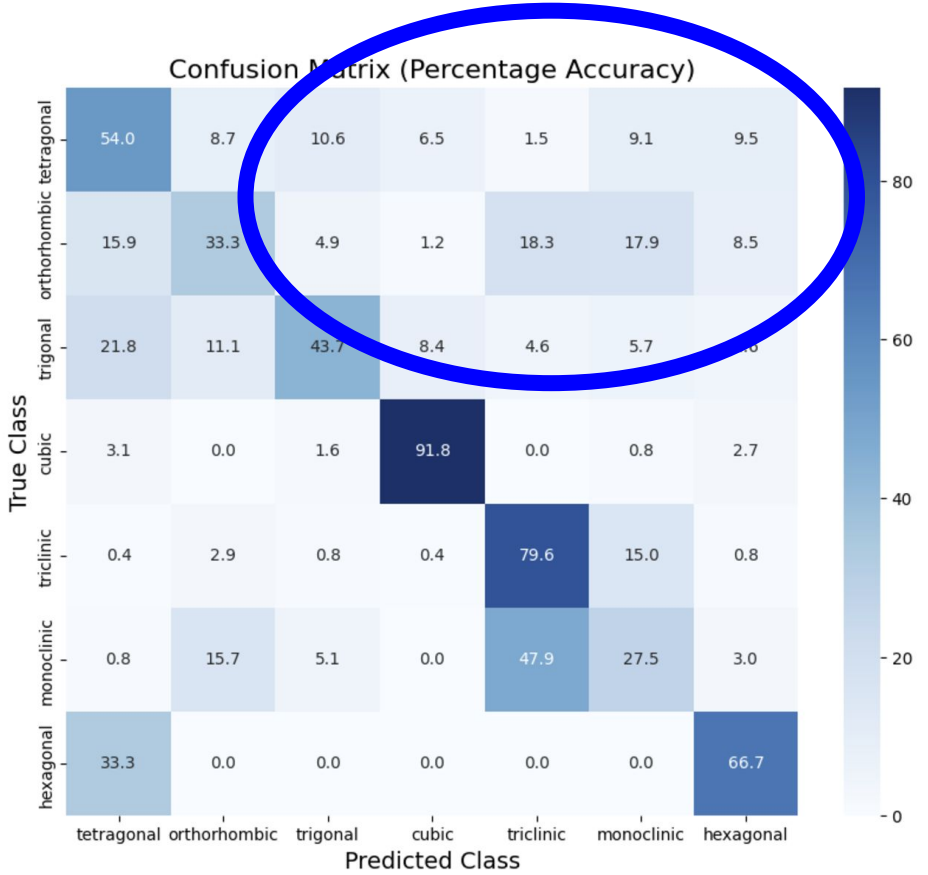
Cross Entropy 64.65% vs. Focal Loss 54.88%

CNN-LSTM, Peak Data

CE/ FL - Confusion Matrix

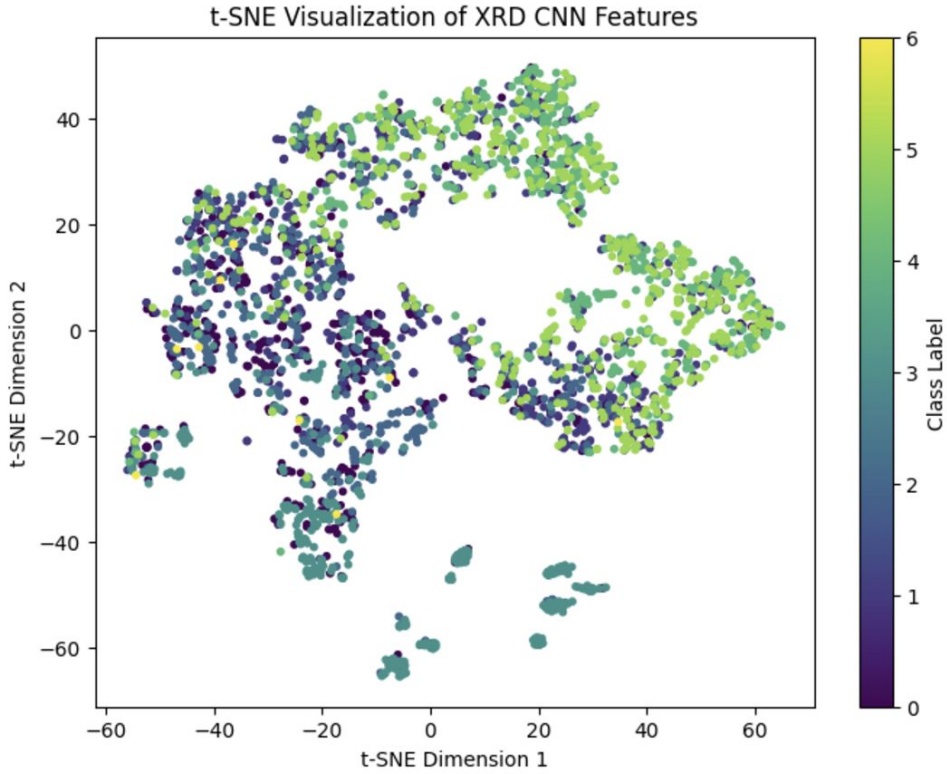


Cross Entropy

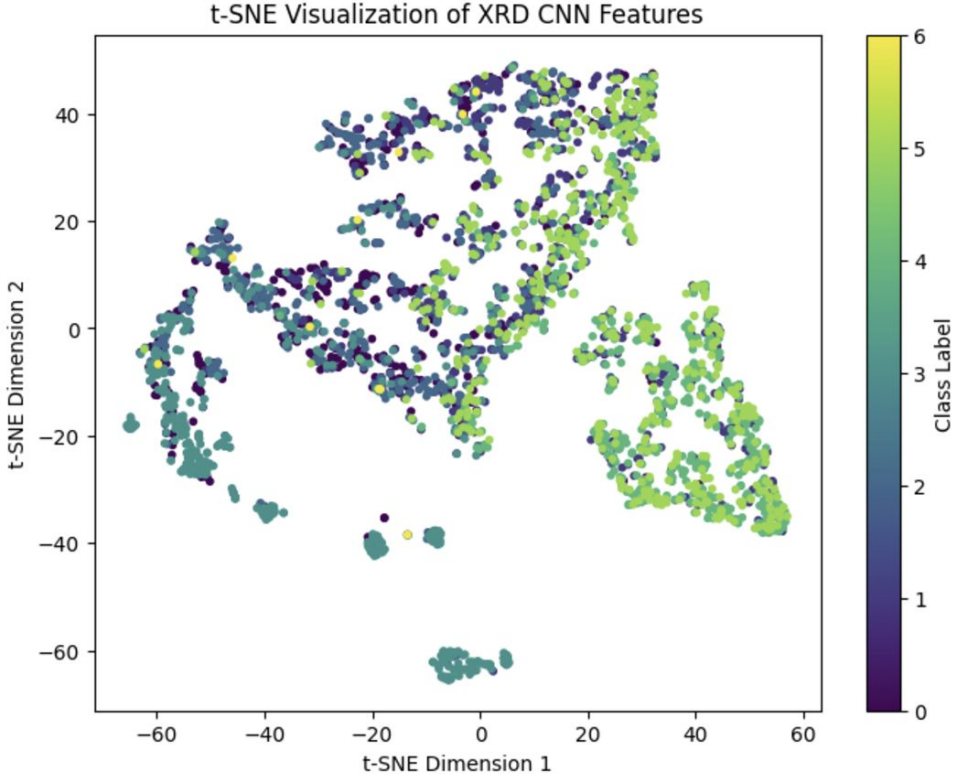


Focal Loss

CE/ FL - TSNE

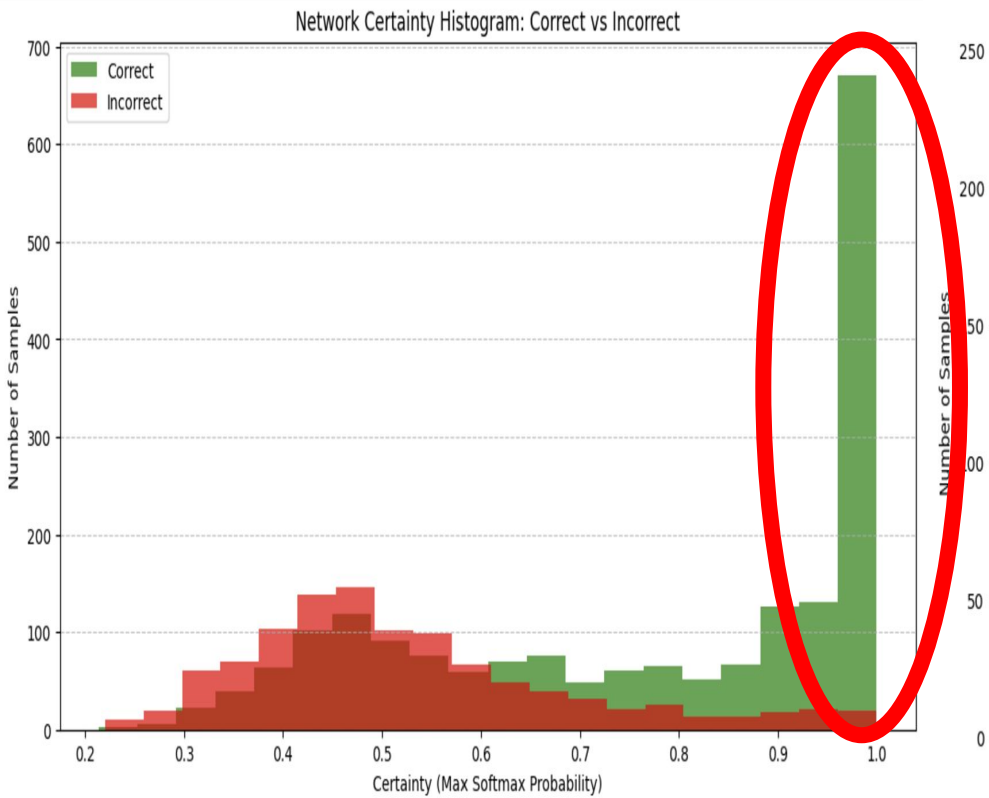


Cross Entropy

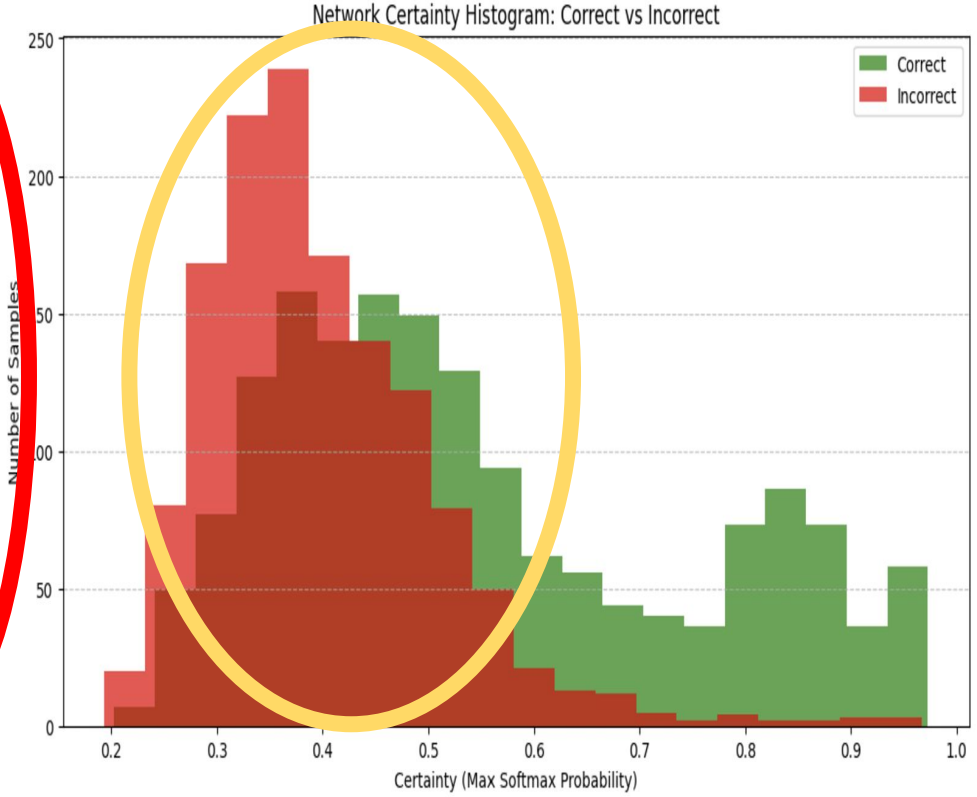


Focal Loss

CE/ FL - Network Certainty



Cross Entropy

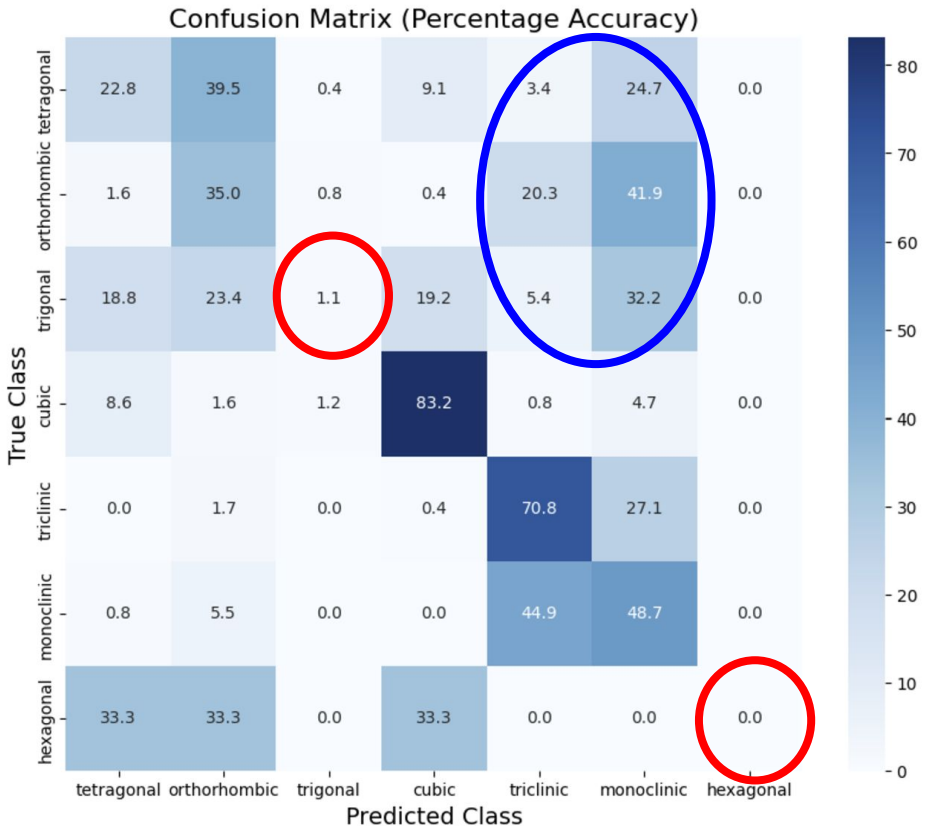


Focal Loss

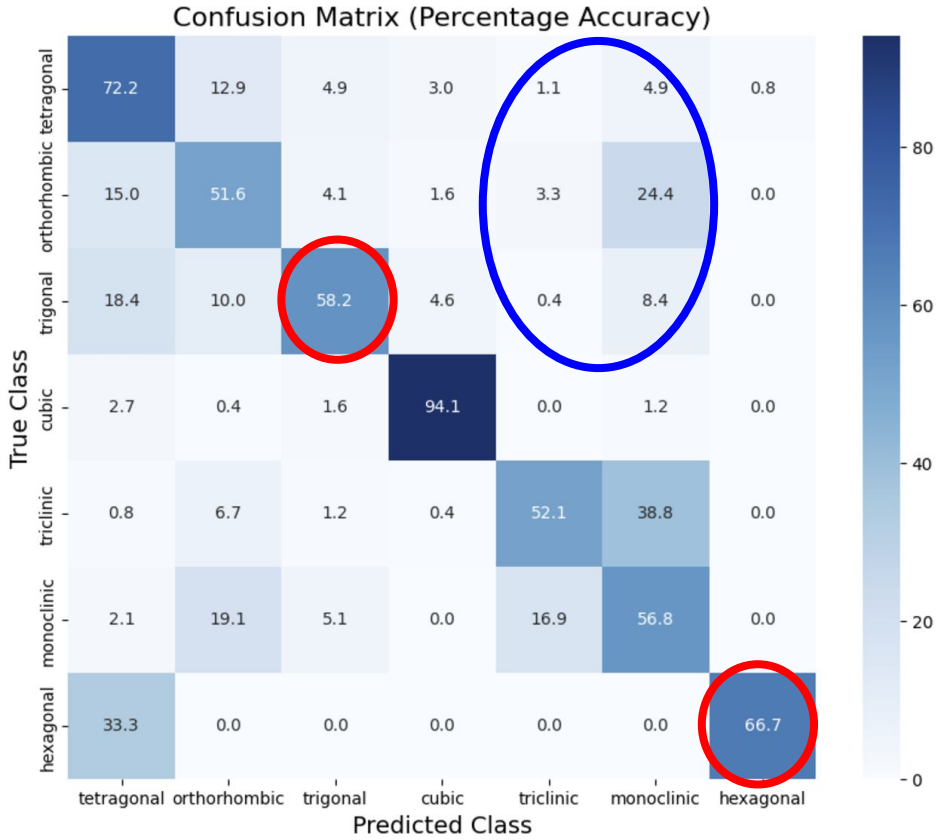
Comparison:
Small Model 43.19% vs. Big Model 64.65%

Cross Entropy, Peak Data

Peak Data - Confusion Matrix

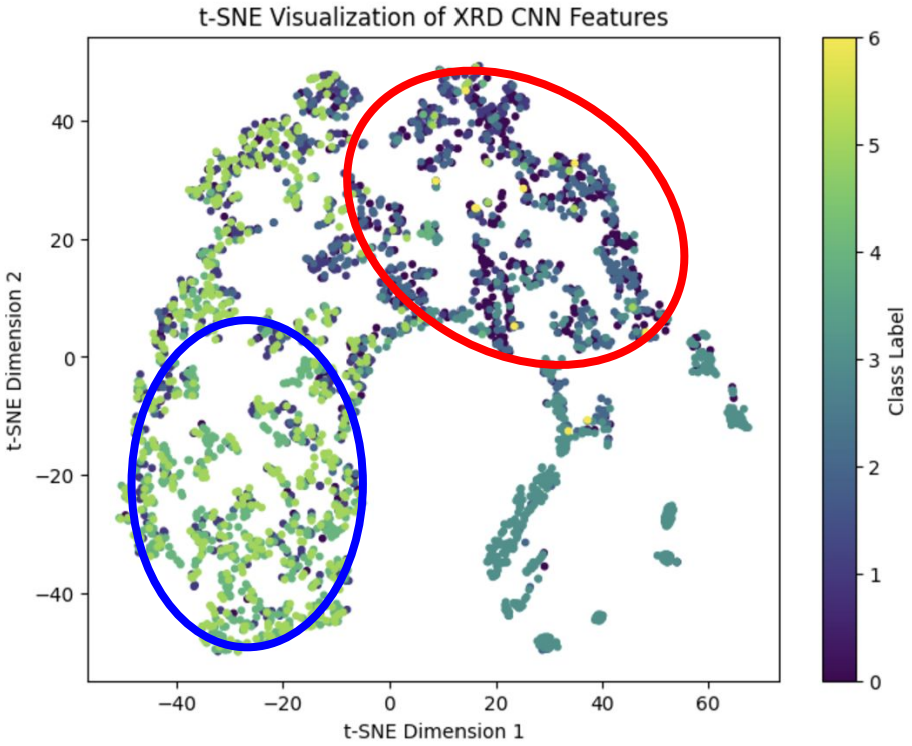


sCNN-LSTM

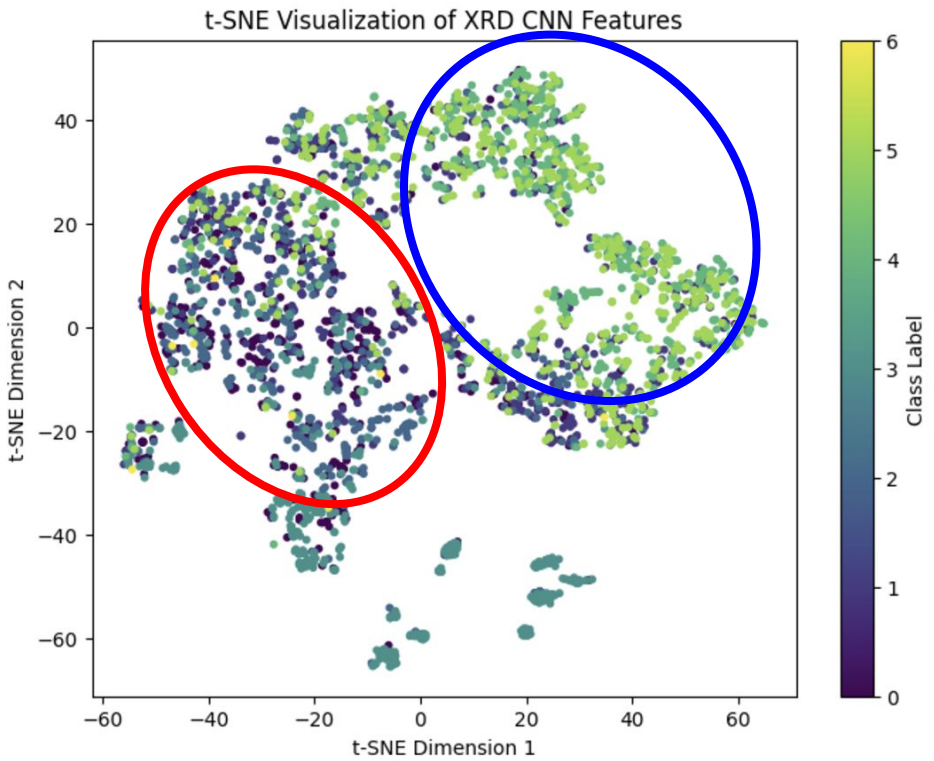


CNN-LSTM

Peak Data - TSNE

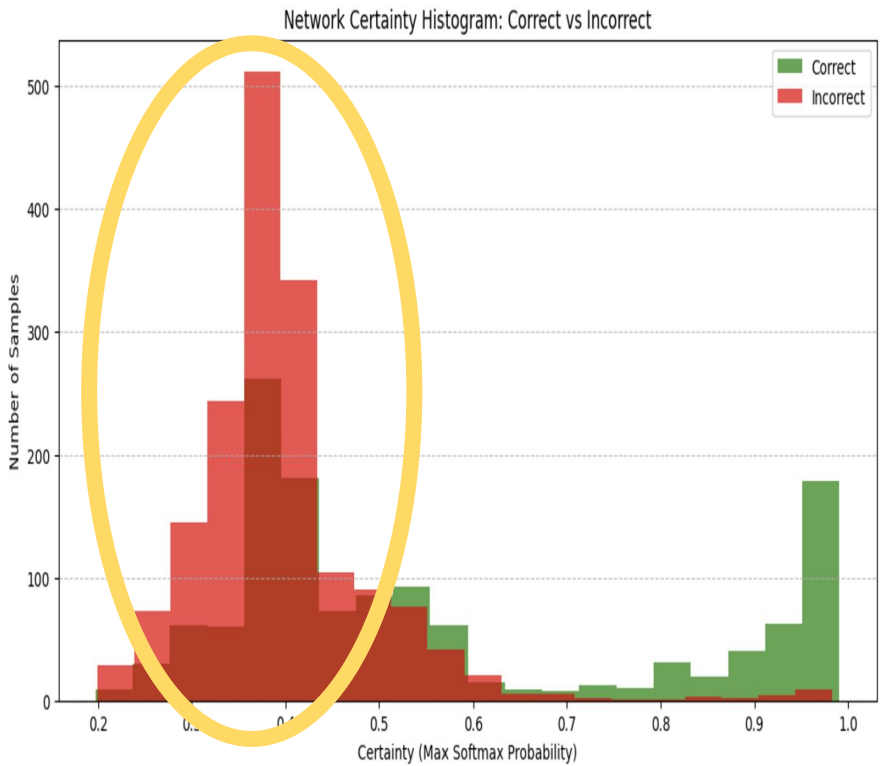


sCNN-LSTM

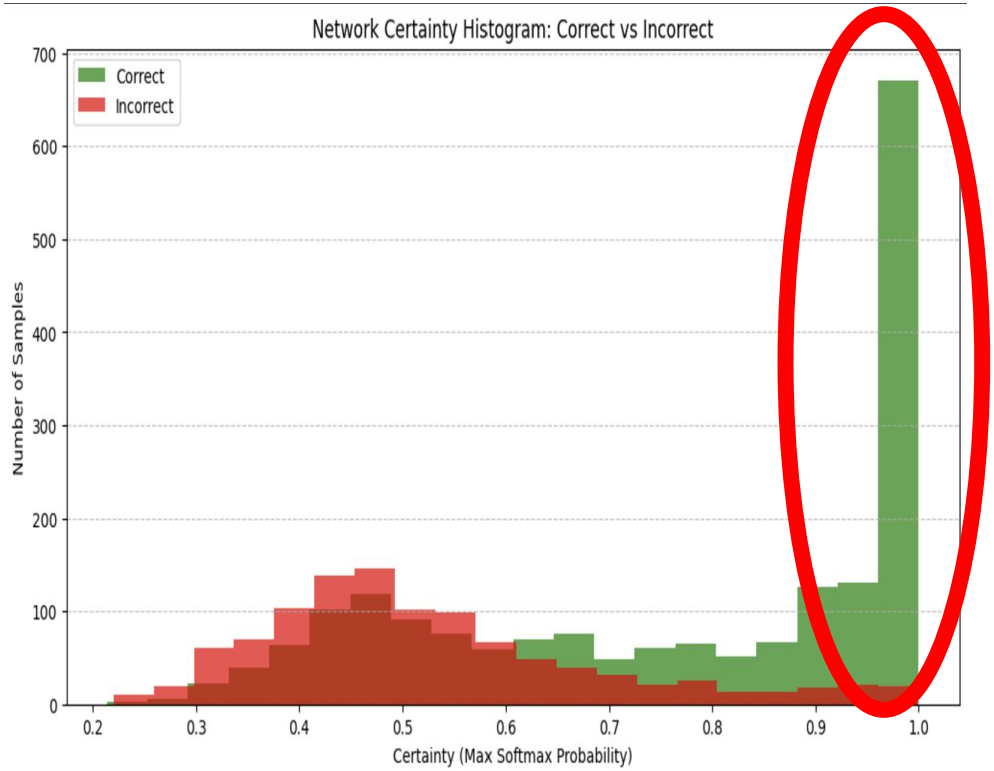


CNN-LSTM

Peak Data - Network Certainty



sCNN-LSTM



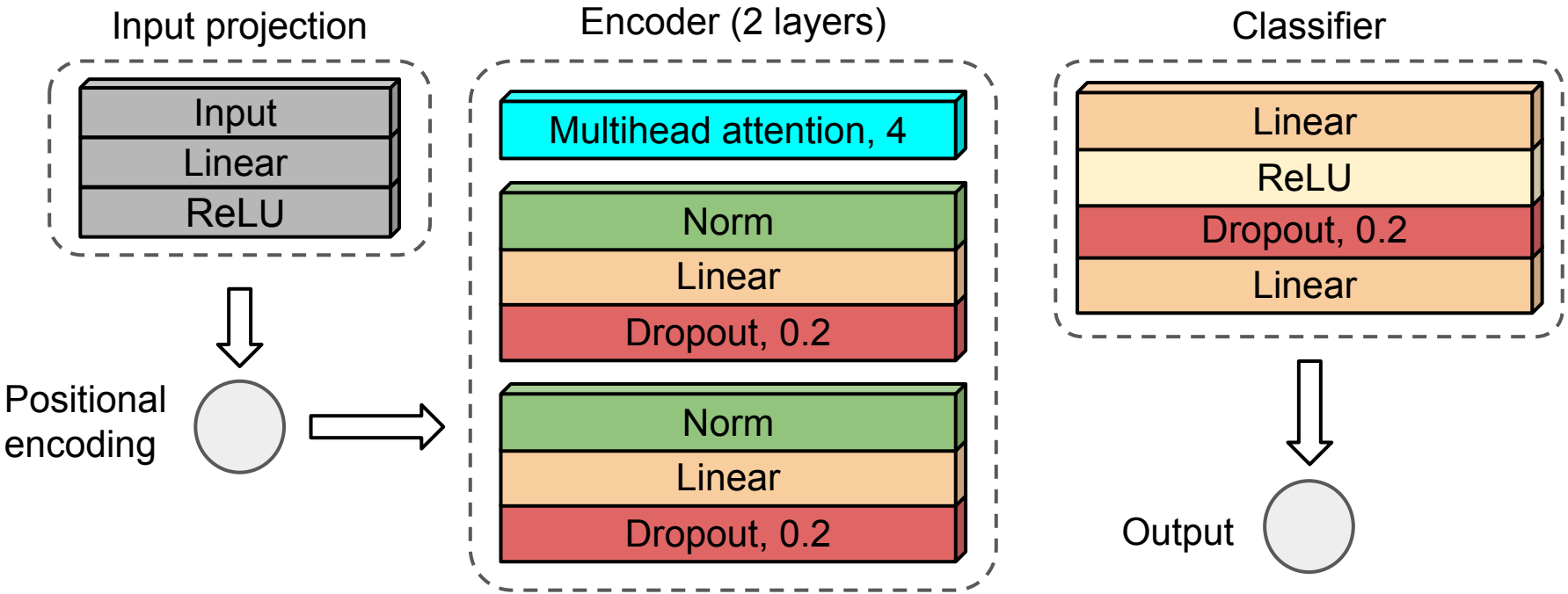
CNN-LSTM

Small Conclusion

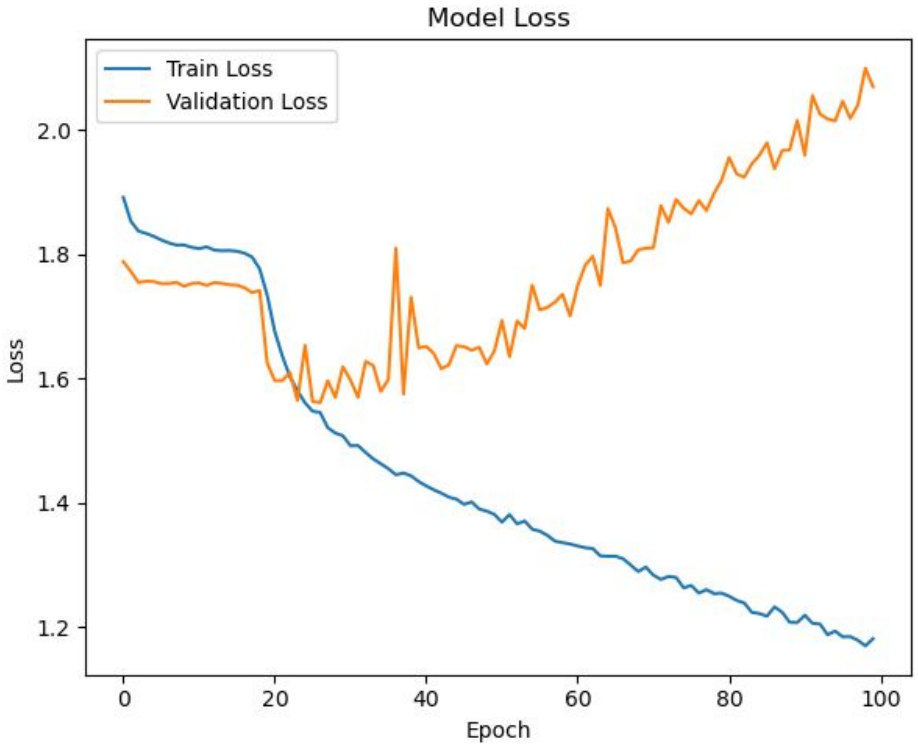
1. sCNN-LSTM works better on big dataset.
2. Focal loss does well on small samples but sacrifices overall performance.
3. CNN-LSTM (big model) works better on peak dataset than sCNN-LSTM.

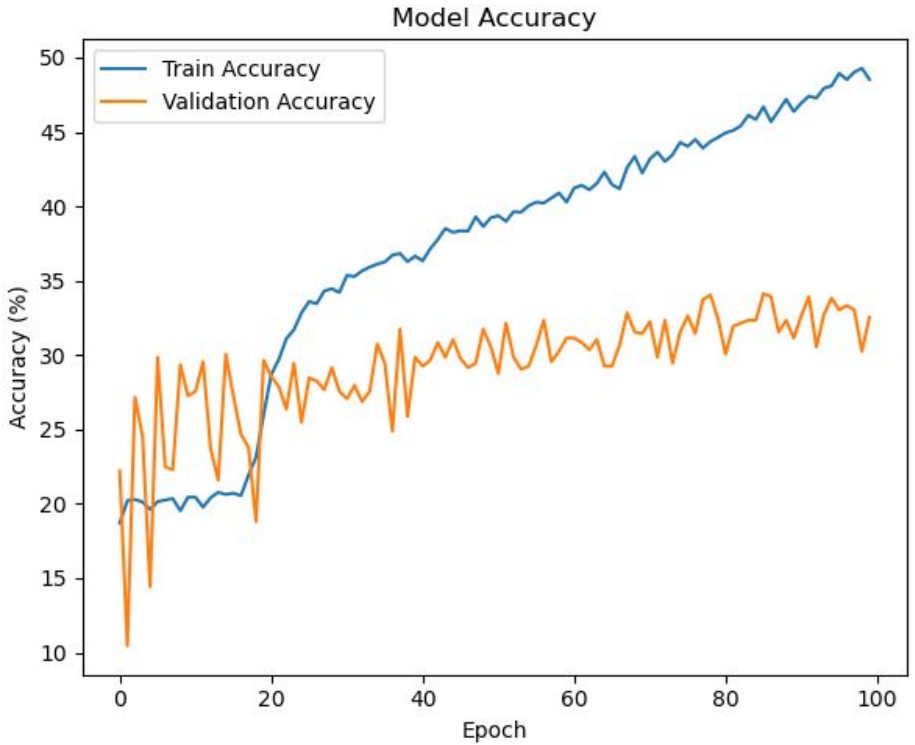
5. Transformer

Architecture

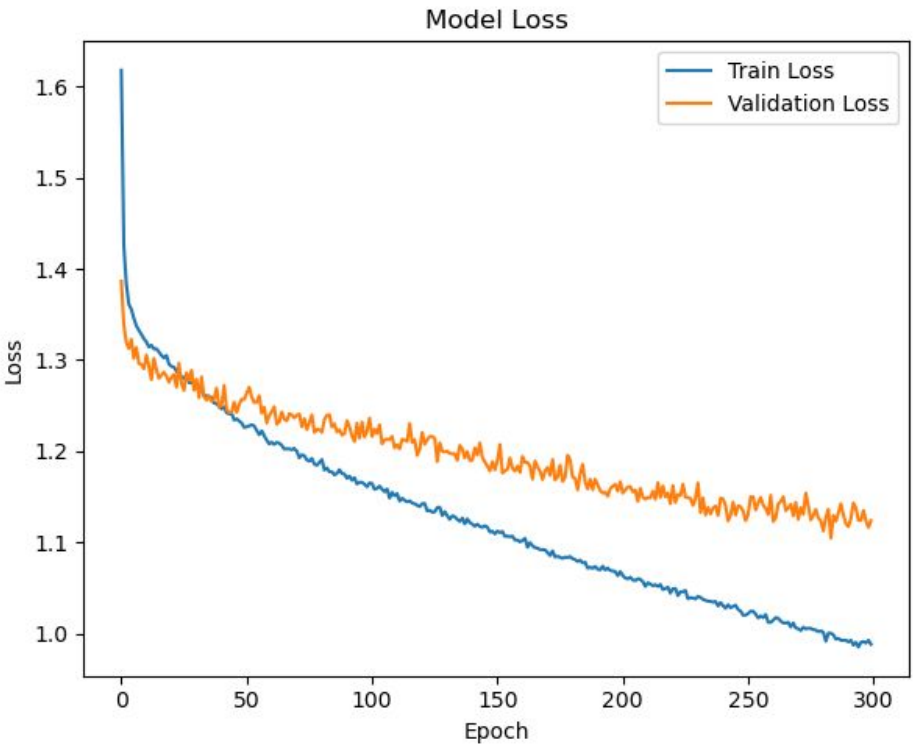


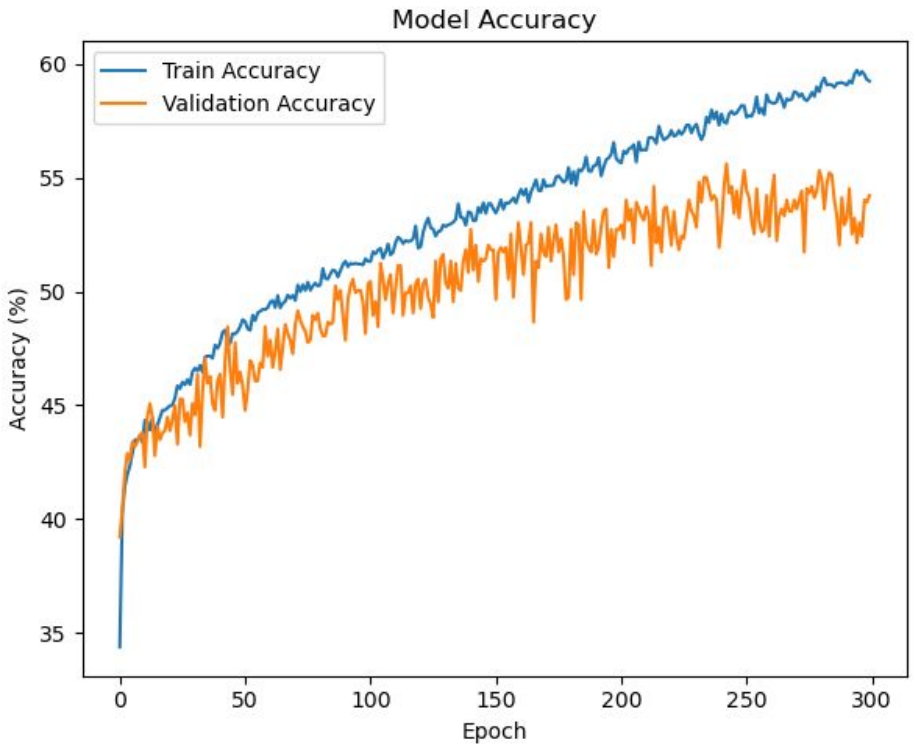
Full spectrum



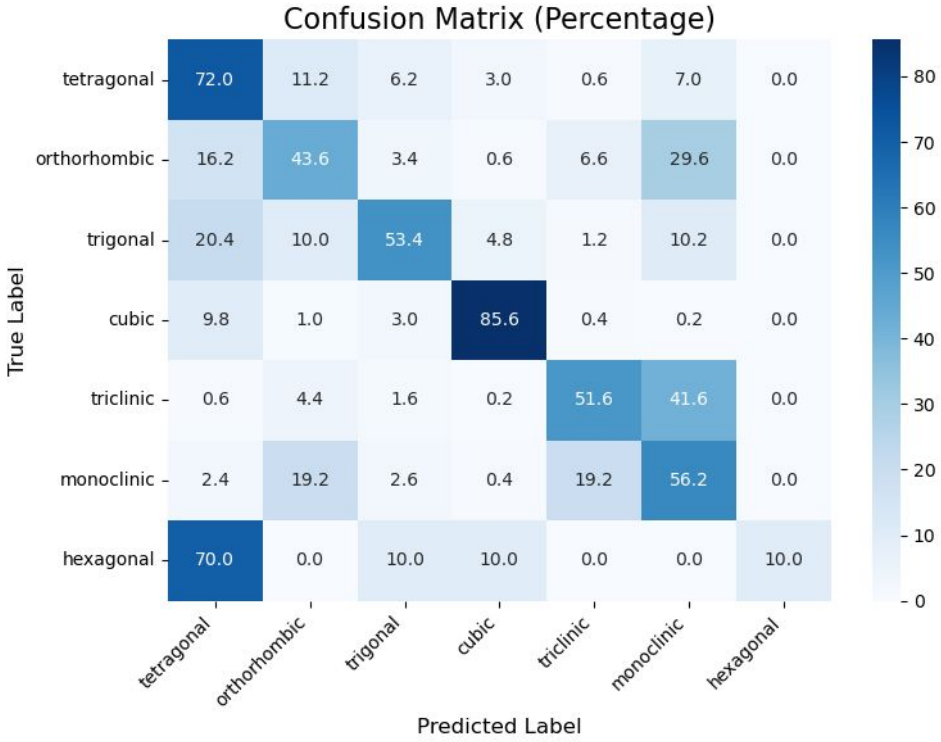


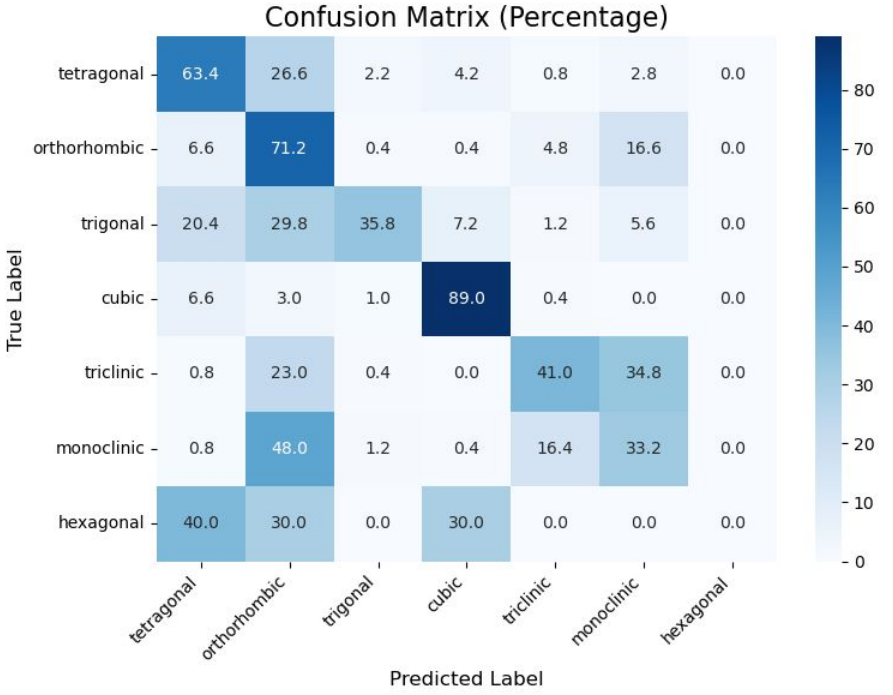
Peaks only


Model Loss
Train Loss
Validation Loss
Loss
Epoch
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0
50
100
150
200
250
300

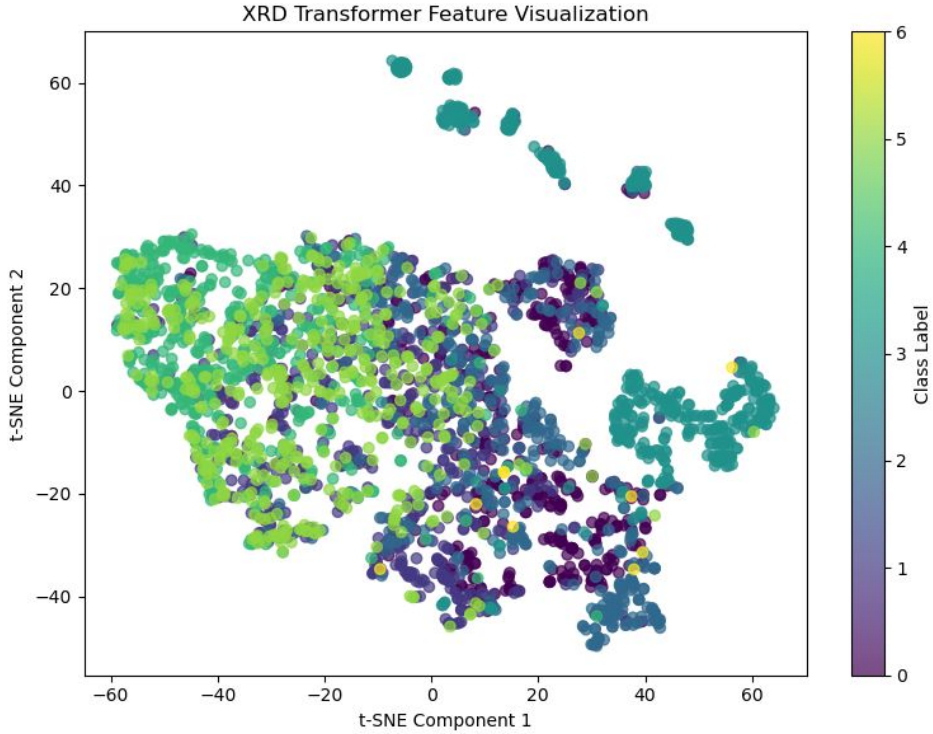

Model Accuracy
Train Accuracy
Validation Accuracy
Accuracy (%)
Epoch
60
55
50
45
40
35
0
50
100
150
200
250
300

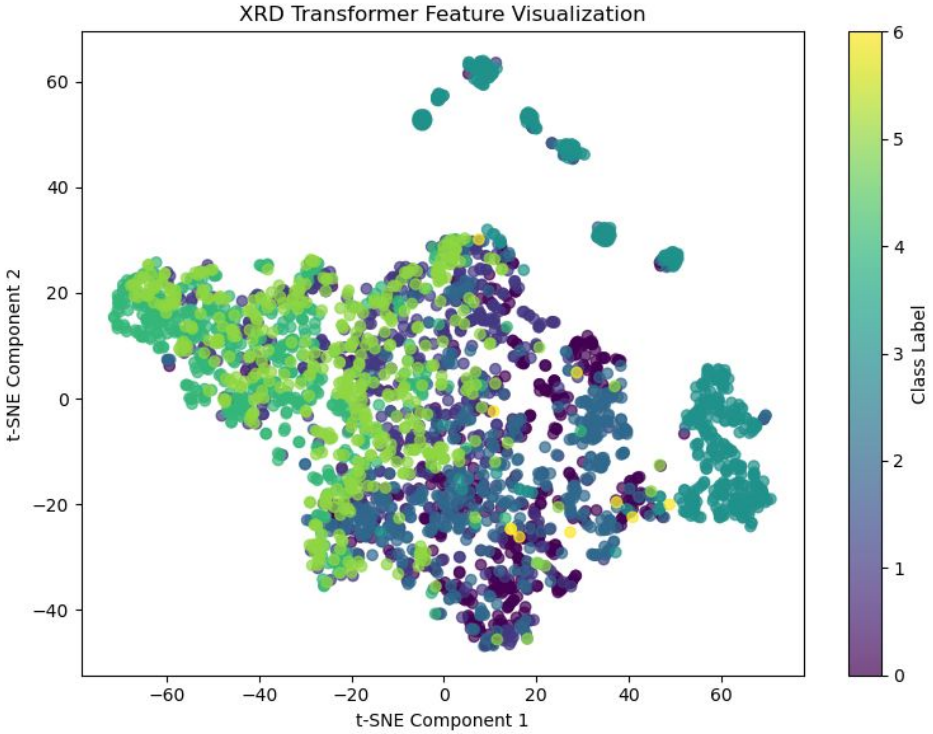
Confusion matrices (best acc: 60%)



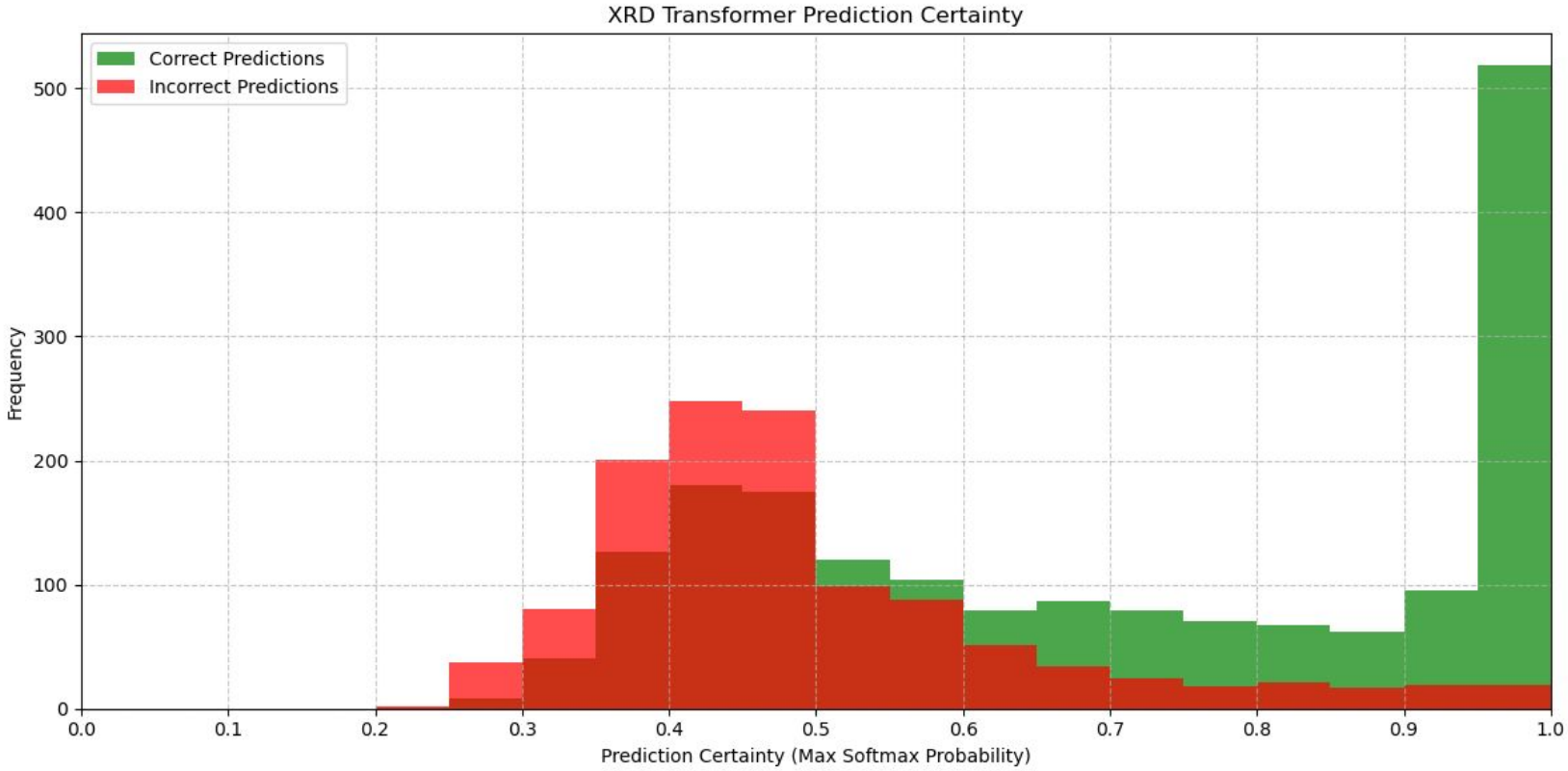


Transformer t-SNE





Certainty


XRD Transformer Prediction Certainty
Correct Predictions
Incorrect Predictions
Frequency
0
100
200
300
400
500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Prediction Certainty (Max Softmax Probability)

Summary / Final remarks

Our models

Model	Full spectrum (%)	Peaks only (%)
Dense (150 MB)	35	56
CNN+Dense (160 MB)		60
CNN-LSTM with CE/FL (1 MB)		64/54
sCNN-LSTM with CE/FL (75 KB)	52/50	43/42
Extreme Randomized Tree (90 MB)		67
Transformer (460 KB)	33	60

Future improvements

Training-wise

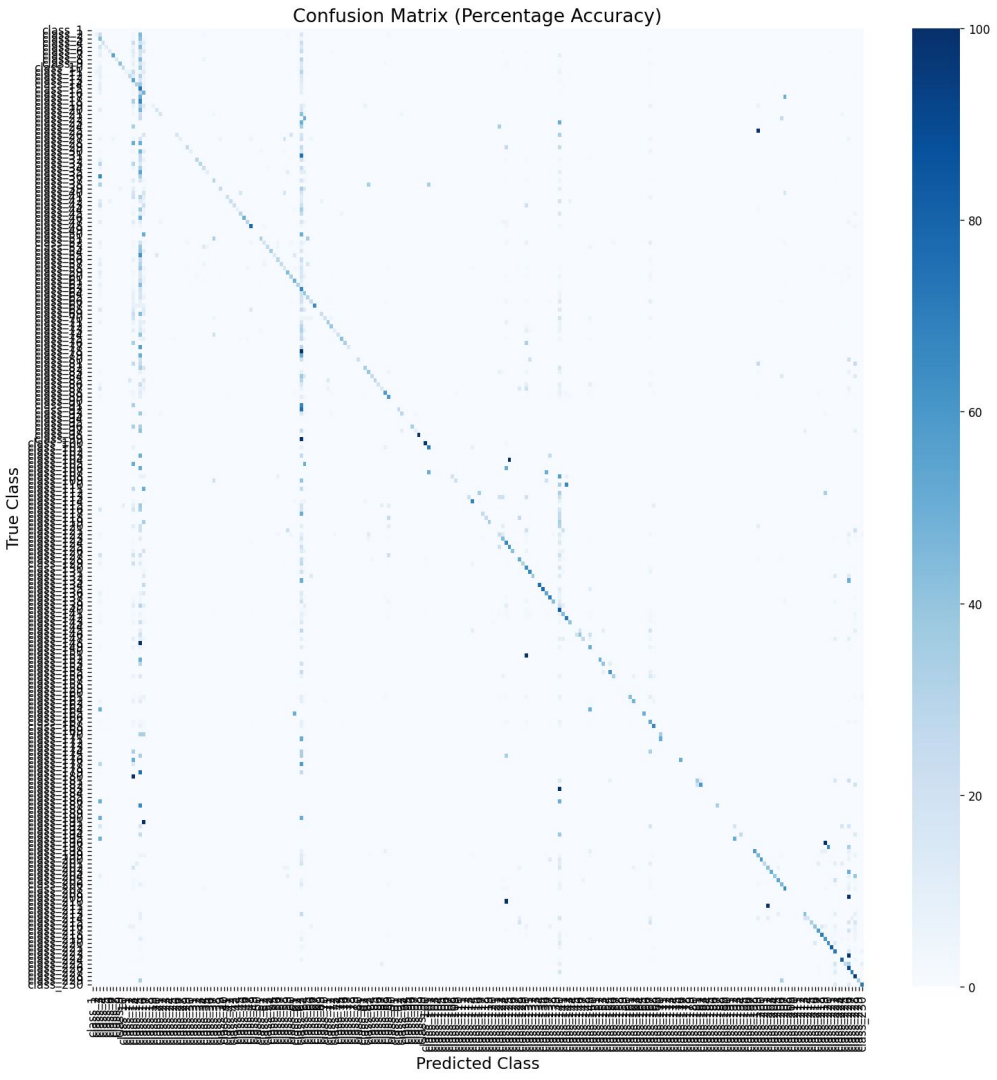
- Extract other physical features
- Try new models (e.g., RNN)

Data-wise

- Larger datasets
- Filter noise
- Normalized spectrum

Supplementa

We have trained t d Extra Trees Classifier.



References

[1] PhysRevB.99.245120

[2] Nature Scientific Reports “Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach”