Prediction of Lattice Structure of Perovskites from physical properties using ML

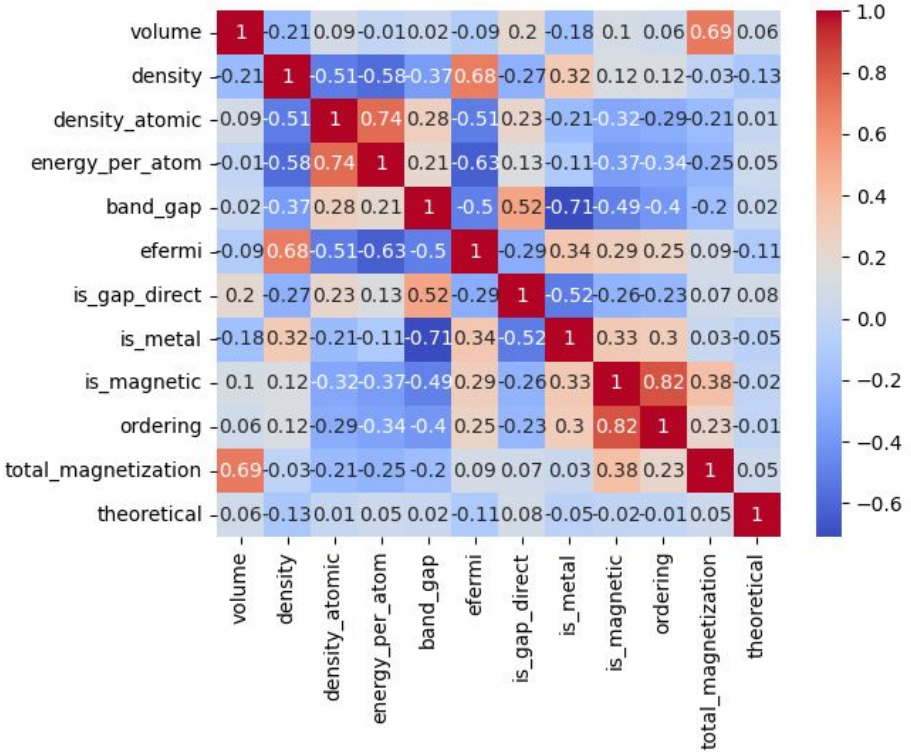
- Group 22: Sunil S Raja (SPS, Batch 20), Kaling Vikram Singh (SPS, Batch 20).
  Guide: Dr Subhankar Mishra

Comparison of contribution with Previous models/Work

Previous Models	Our Model
Dataset had only pure oxide perovskites with no double perovskites etc	Data set contains oxides and halide perovskites with double perovskites etc
Data set had only pure perovskite(ABX3)	Data set has mixed perovskite(ABC2X6)
Included Tolerance and radius as a feature	Did not include Tolerance radius as a feature

Data Sets

- Most of the additional data was downloaded from Materials Project API with MPRester and robocrys modules.
  (https://materialsproject.org/api)
- We notice that most of the data is not correlated to each other.
- Oversampling was done to make the dataset balanced.



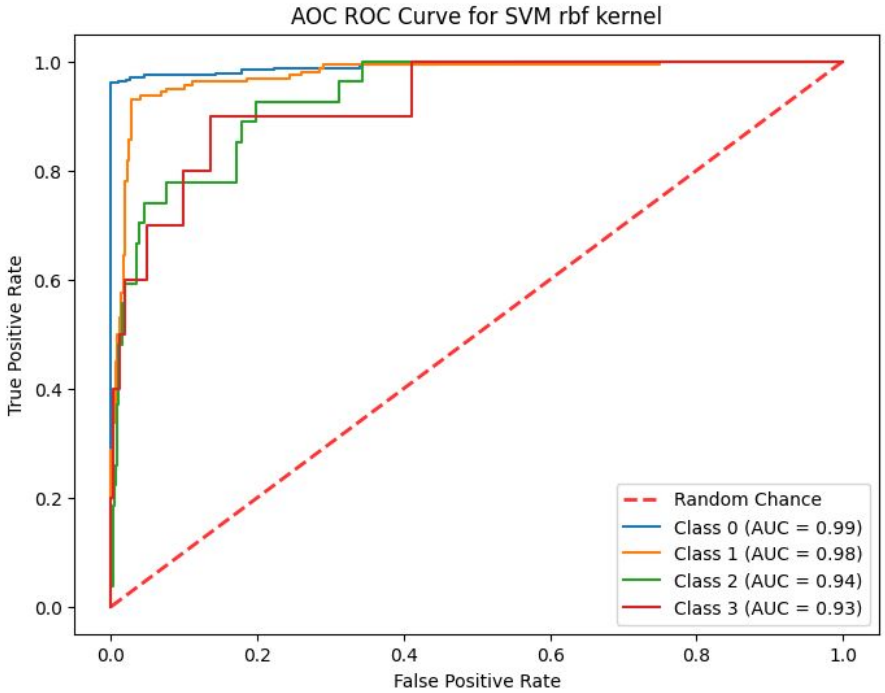
Features

- For Midway we included features such as valency A, valency B, radius A and Radius B, as only oxides were in the dataset. This however did not work when the dataset has both oxides and hallides as ABO3 and ABF3 both will have the same value of the 4 features discussed above.

- We found that volume of the lattice is a better feature than volume per atom. As this also could potentially help differentiate between if it is a ABX3 perovskite or AB3X6 perovskite

Models and Hyper-parameter tuning

- Multiple Multiclass classification modules were implemented.
- The data splitting was done with sklearn.
- The hyper parameter tuning was done using Gridsearch or by observing the ROC curve.



Changes after Midway

Midway	Final
No of features =10	No of features =12
Features such as radius and valency of atoms are considered.	Features such as magnetic ordering and fermi energy.
Fit takes less time	Fit takes more time
Less Models	More Models(Random Forests,KNN)

Midway Progress

- The Baseline was successfully built and implemented for oxide perovskites.
- SVM was obtained to be most accurate classifier with 92.32% accuracy.

Accuracy vs Model	
Model	Accuracy
SVM	91.33%
Weighted SVM	92.03%
Light GBM	94.32%
XGBoost	94.13%

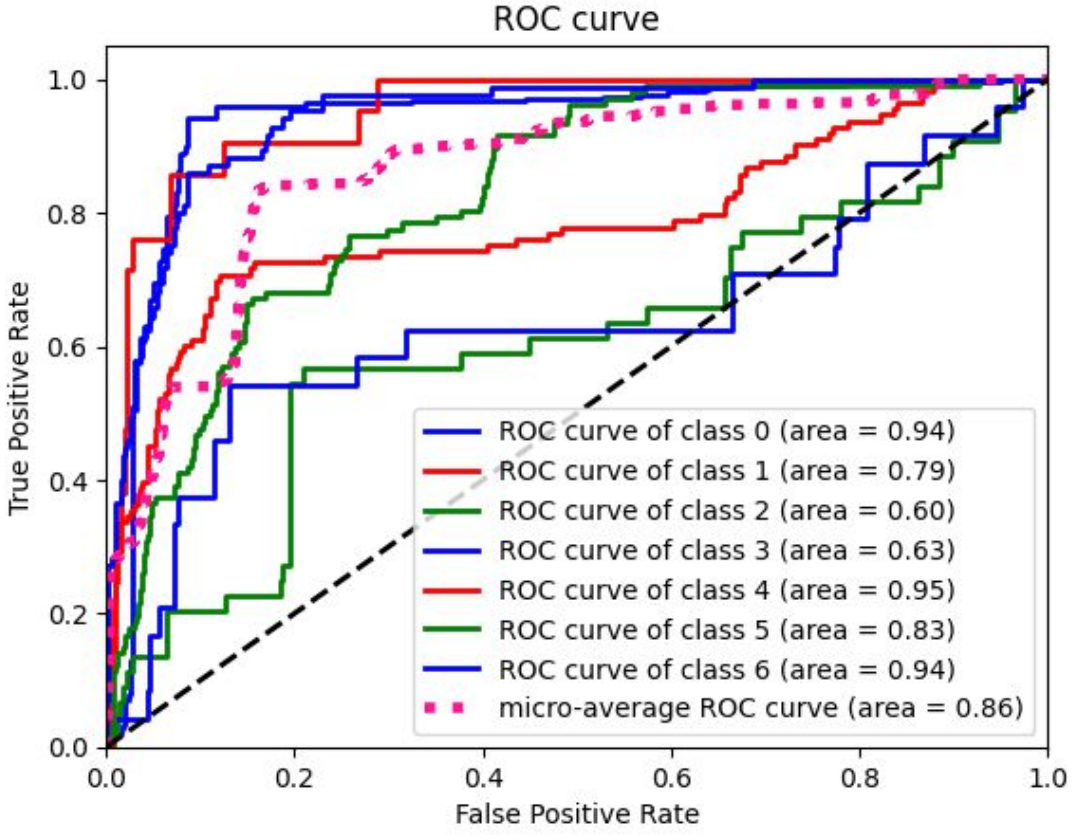
“Future” Goal

- Implementation of the model for halide perovskites

SVM

- Kernel: RBF
- C(Penalty parameter): 19

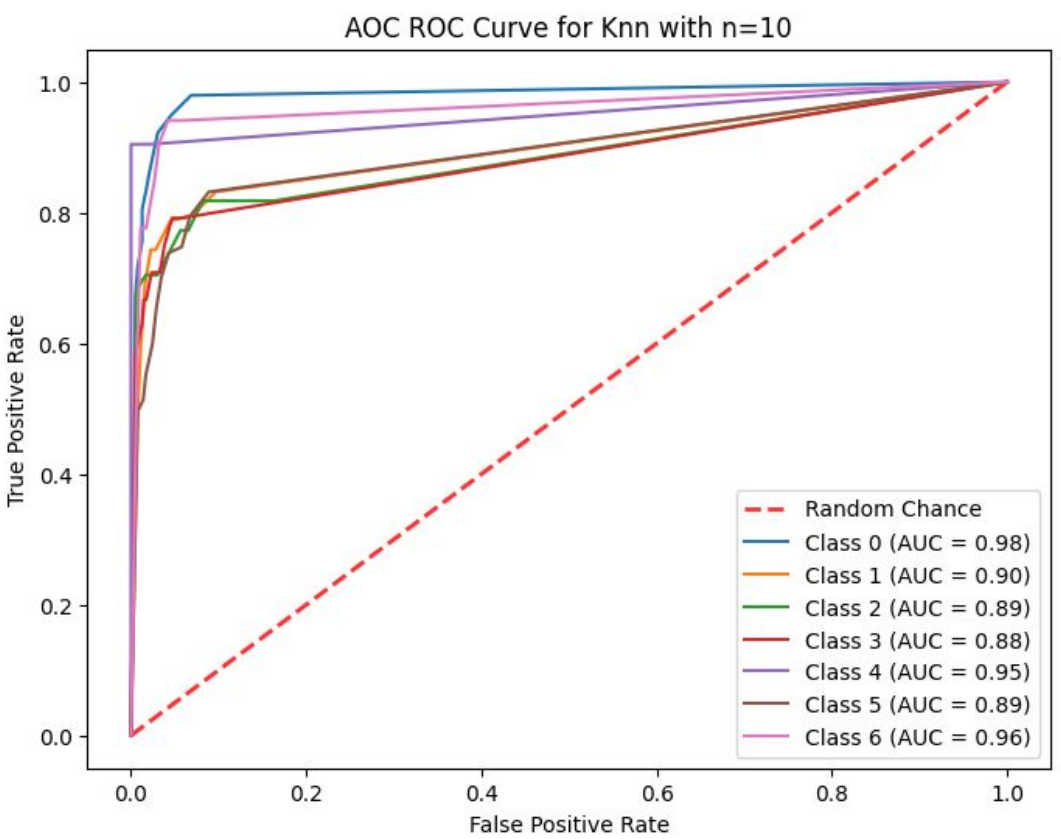
- Results
  - AUC ROC: 0.86
  - Accuracy: 55.26%



KNN

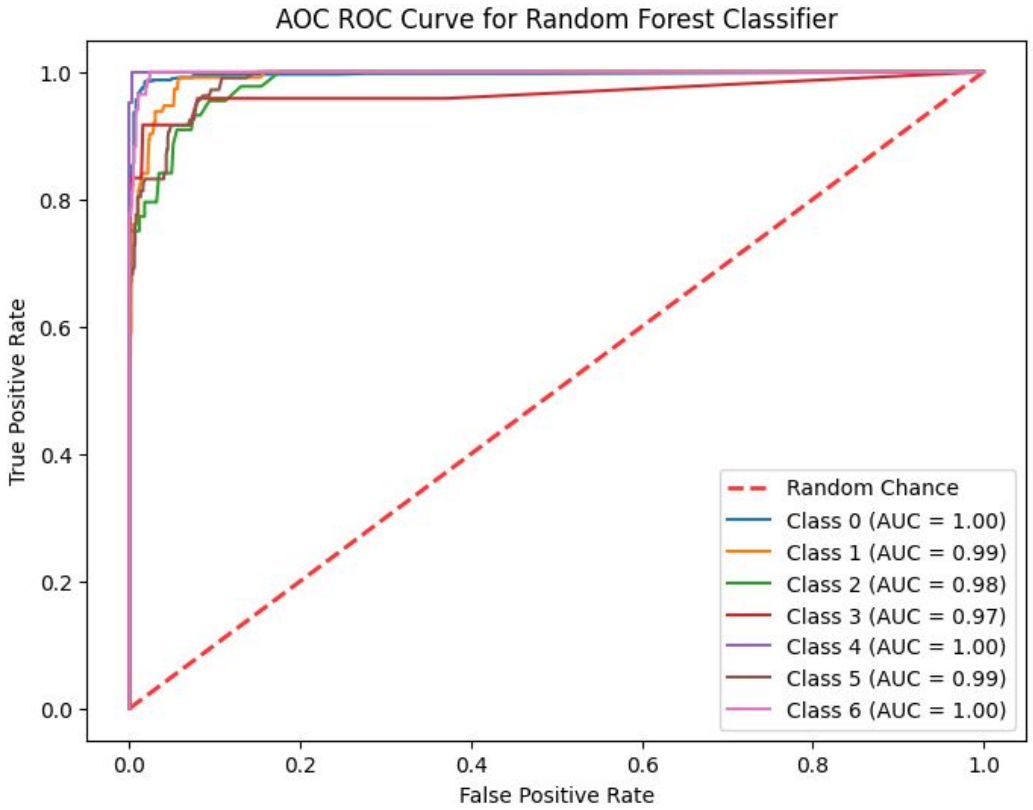
- `N neighbours= 10`

- Results
  - AUC ROC: 0.92
  - Accuracy: 82%



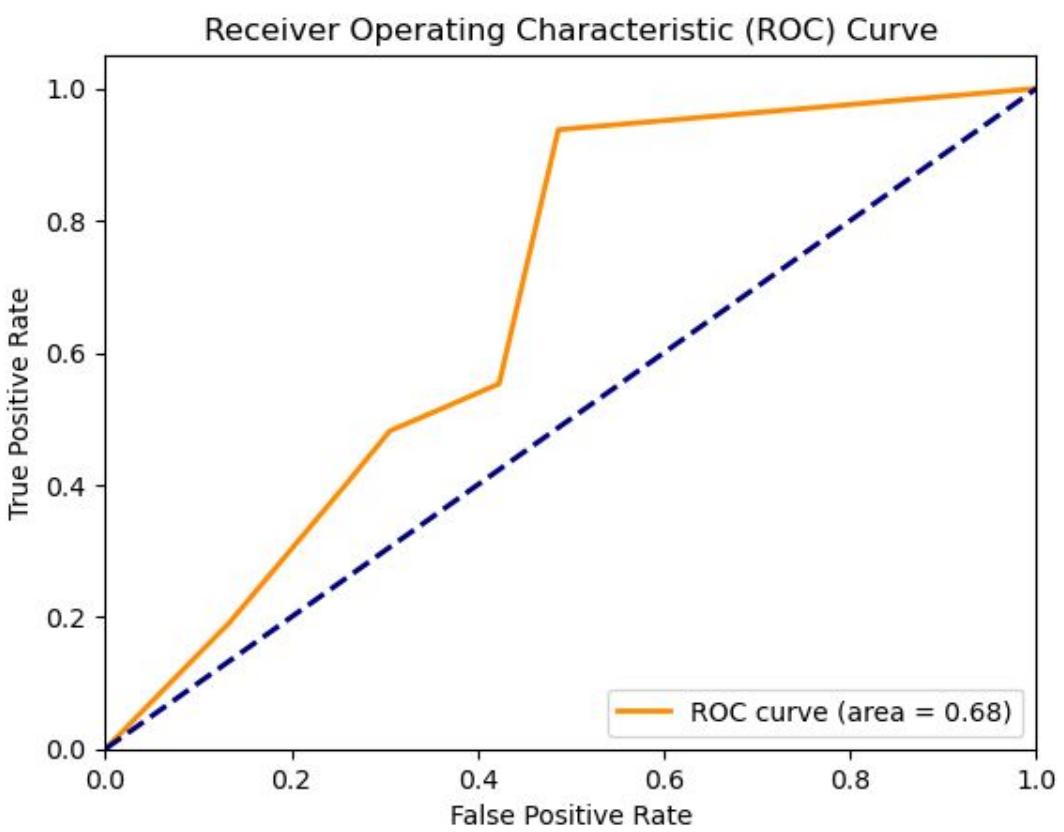
Random Forests

- `n_estimators: 250`
- `criterion: entropy`
- `max_depth: 100`
- `min_samples_split: 2`
- Results
  - AUC ROC: 0.98
  - Accuracy: 94.4%



XGBoost

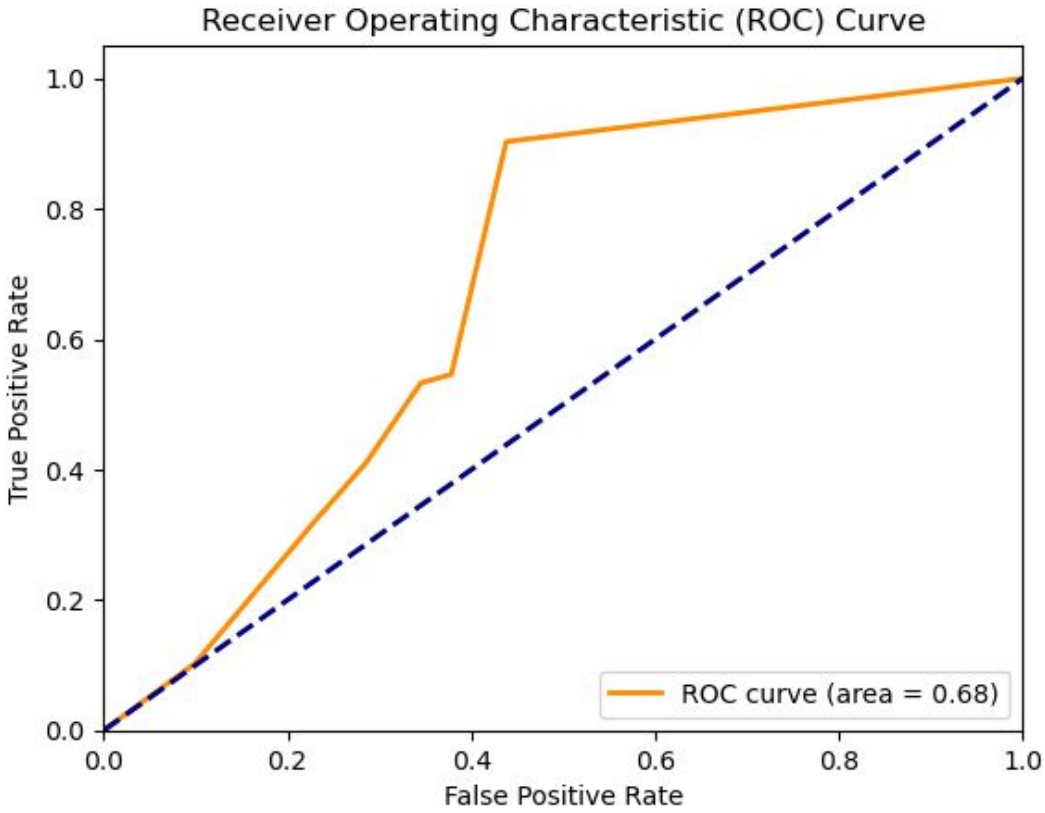
- `epochs: 1`
- `learning_rate: 0.7`
- `tree_height: 4`
- `loss: logarithmic`
- Results
  - AUC ROC: 0.68
  - Accuracy: 60.86%



LGBM

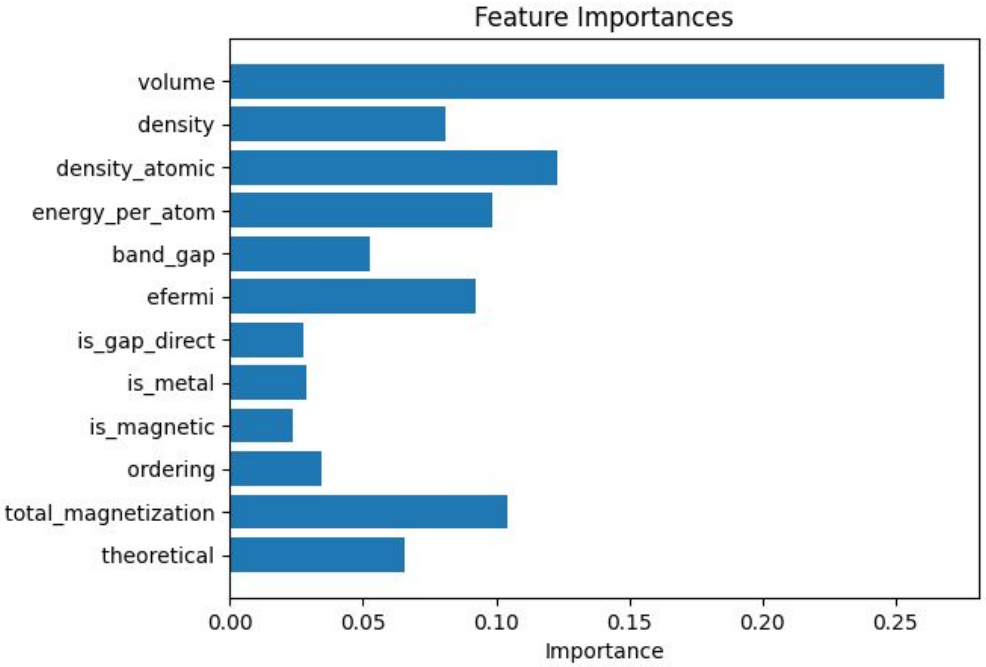
- `learning_rate: 0.01`
- `tree_height: 3`
- `loss: logarithmic`

- Results
  - AUC ROC: 0.68
  - Accuracy: 60.86%



Areas of possible improvement

- We have multiple features with binary values, after SMOTE we observe some non binary values. This will lead to errors
- Further improvement in accuracy
- Lot more time required to fit the model for perovskite halide and oxides model compared to the midway work for perovskite oxides.
- Removing is_magnetic feature should not decrease accuracy, however a 0.3% decrease was observed.



Feature Importance for random forests

Conclusion

- The model was implemented for perovskites oxides and halides, with a similar accuracy in both algorithms(~94%)
- It may be possible to include more types of perovskites like layered perovskites, organic perovskites etc and predict their structure also, this was not done as data was available for open access.

Accuracy vs Model		
Model	Accuracy	AUC-ROC
SVM	55.26%	0.86
Random Forests	94.4%	0.98
Light GBM	60.21%	0.68
XGBoost	60.86%	0.68
KNN	82%	0.92

Path to paper publication

- Computational Material Sciences
- Journal of material informatics
- Chemistry of materials