DTSA 5509 Supervised Learning Final Project

Semiconductor Wafer Map Failure Pattern Classification

Daisaku Ikoma
Sept, 6th, 2023.
Master of Science in Data Science
University of Colorado Boulder

1. Introduction

- Moore's law drives semiconductor advancement through miniaturization and increased wafer diameter.
- Yield improvement is vital in complex semiconductor manufacturing.
- Traditional defect analysis by experienced engineers is limited in advanced semiconductor manufacturing.
- **Machine learning for defect detection, root cause analysis, and yield prediction is gaining importance**, with a focus on a Random Forest, XGBoost and LightGBM for defect classification.

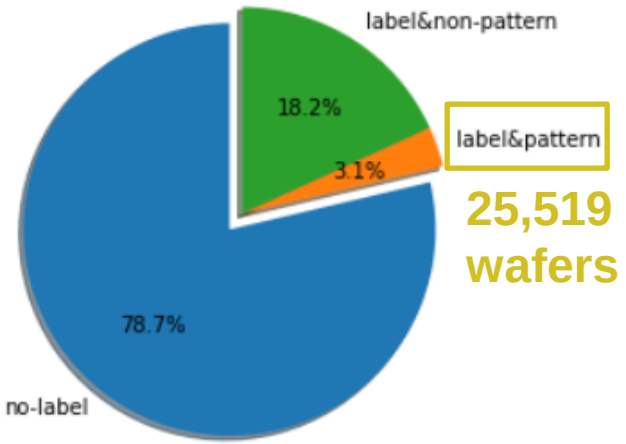
2. Problem statement

- Dataset description:
  - WM-811K dataset provided by MIR lab (http://mirlab.org/dataSet/public/).
  - 811,457 wafer maps collected from 46,393 lots in real-world fabrication.
  - Defects type: Center, Donut, Edge-Loc, Edge-Ring, Loc, Random, Scratch, Near-full, none.

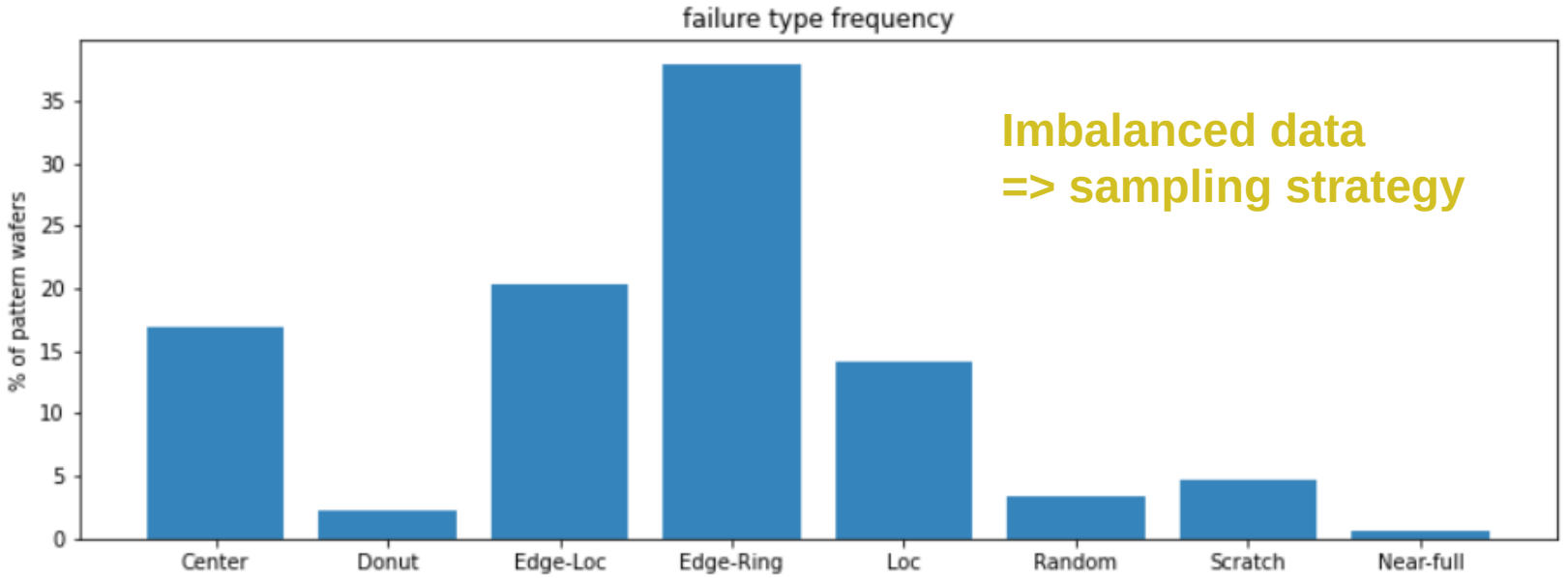
- **Our goal: Identify classes of semiconductor wafer defect maps using machine learning.** In other words, highly accurate **supervised multi-class classification** is achieved.

3. EDA (1/2)

- 172,950 wafers have labels while 78.7% wafers with no labels
- Only 3.1% wafers (25,519 wafers) have real failure patterns
- Focus on **25,519 instances** with high **imbalance distribution**.



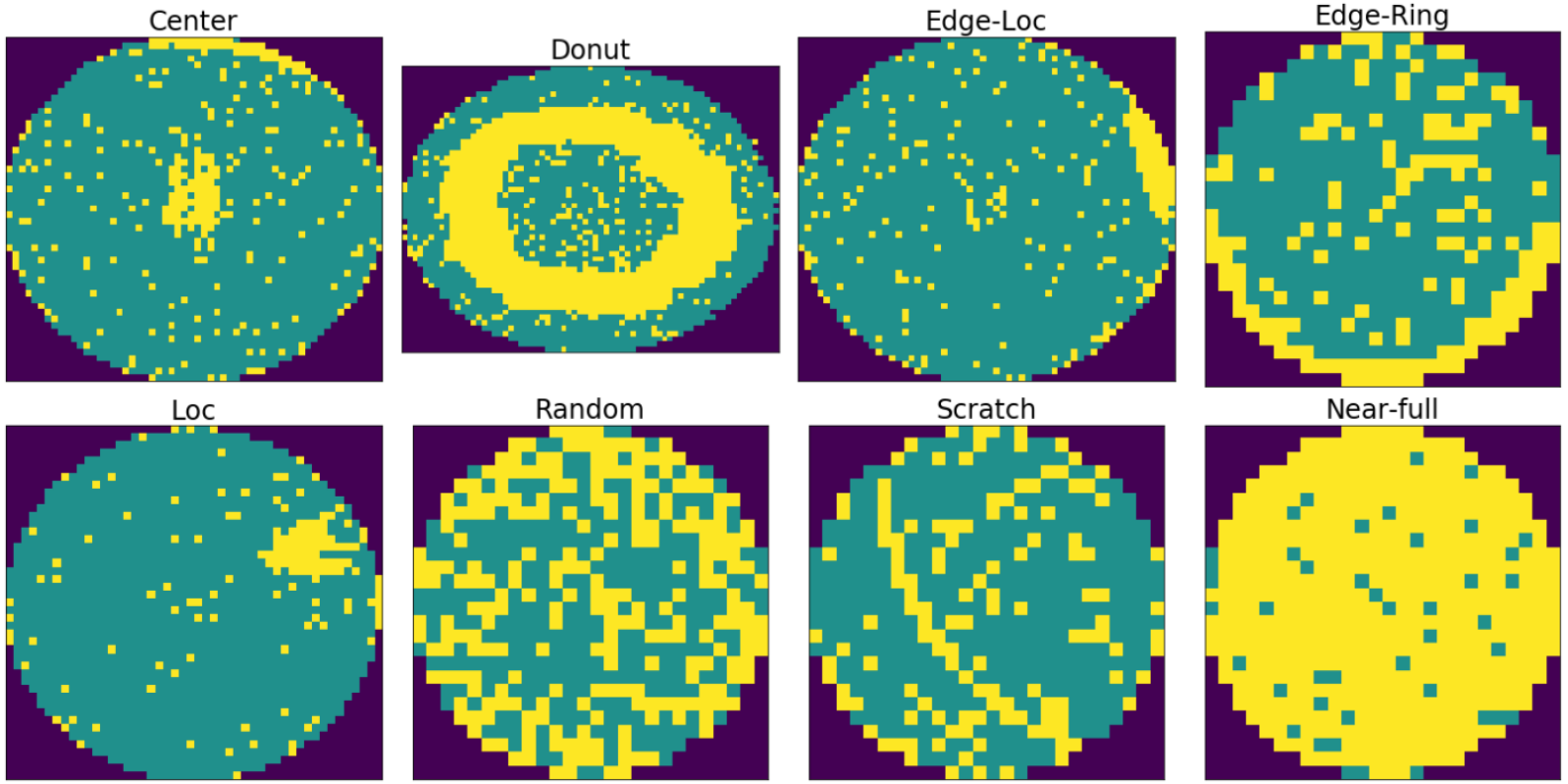
25,519 wafers



Imbalanced data
=> sampling strategy

3. EDA (2/2)

- Typical failure maps (classes) for classification



4. Feature engineering (1/3)

- Density-based feature extraction [1]

Defect density of each region is extracted as 13 features

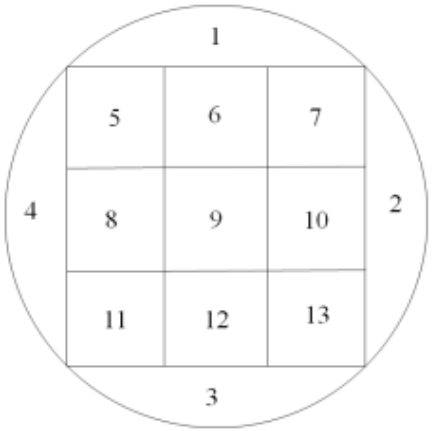
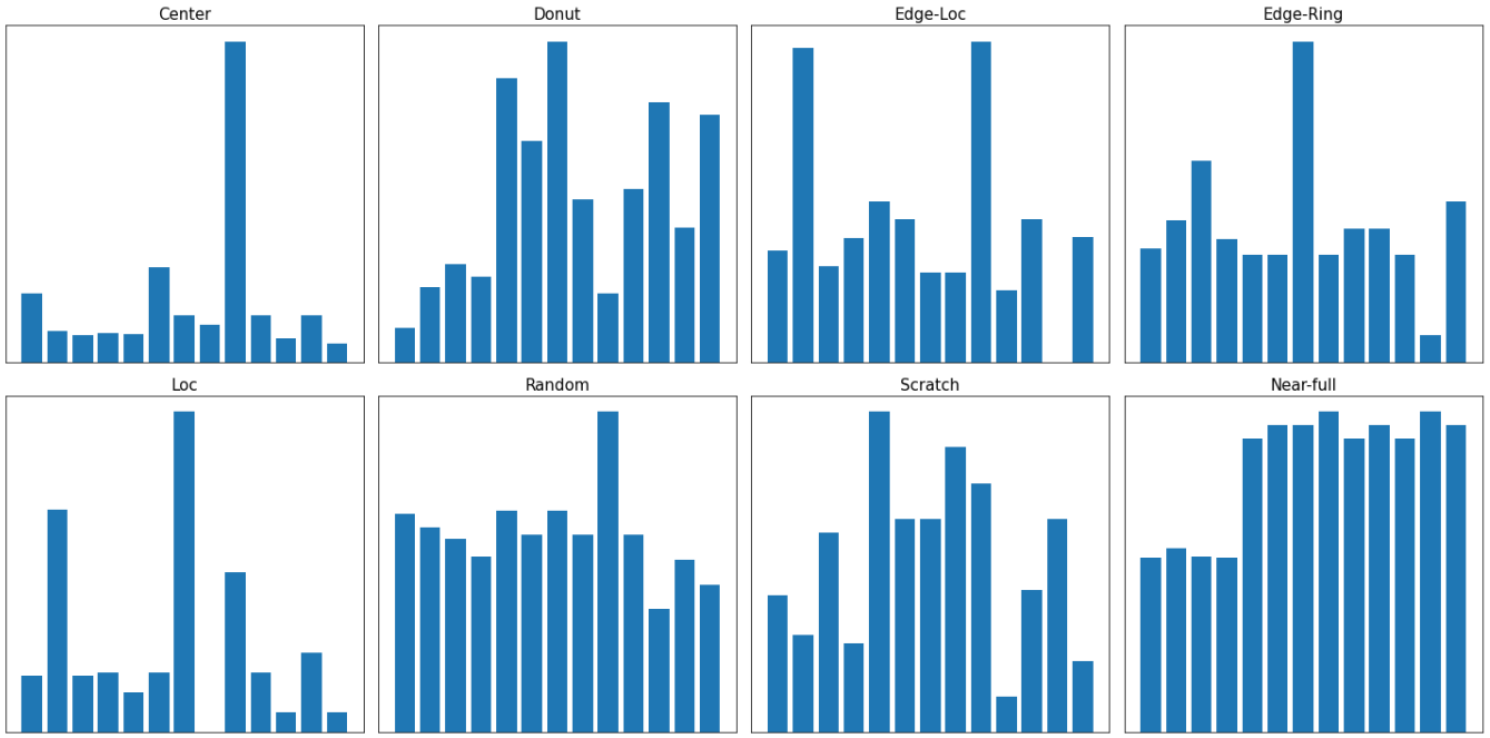


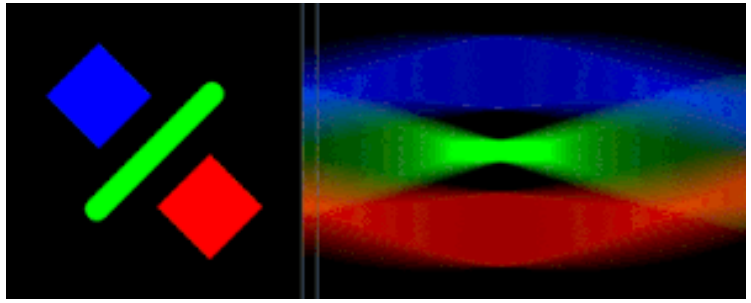
Fig. 4. Density based feature extraction

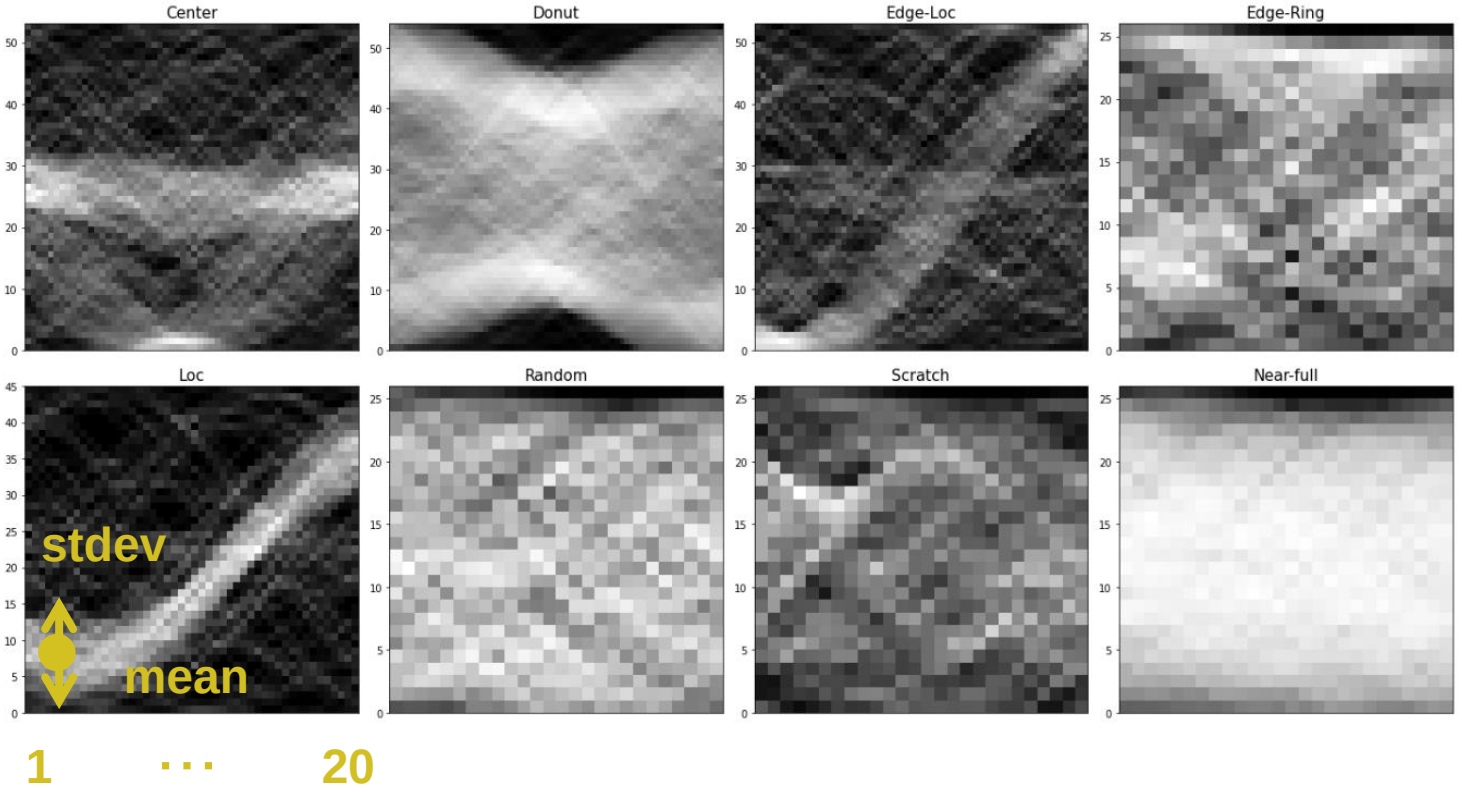


4. Feature engineering (2/3)

- Radon-based feature extraction [3]

Radon-based features are extracted as 20 x 2 (μ, σ) = 40 features





4. Feature engineering (3/3)

- Geometry-based feature extraction [3]

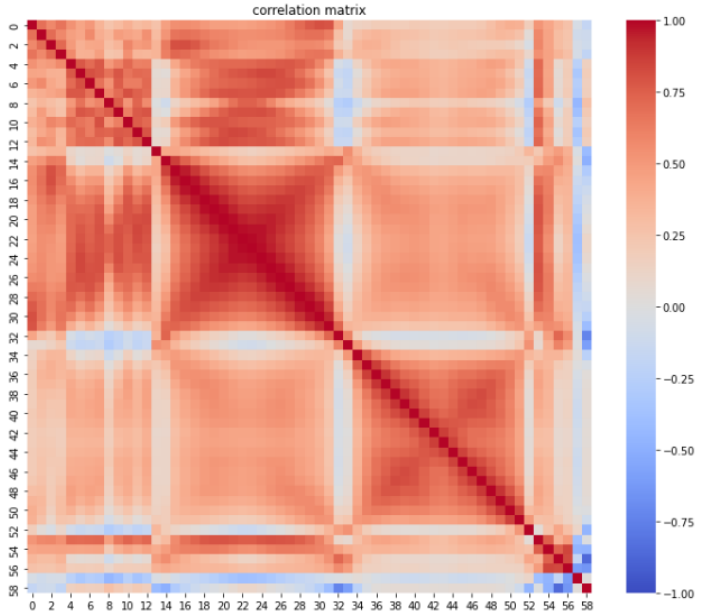
Geometry-based features are extracted as 6 features

Features	Explanation
Area (prop_area)	The area of a region in the image normalized by the total area.
Perimeter (prop_perimeter)	Perimeter of the area in the image, normalized by the diagonal length.
Major axis length (prop_majaxis)	The value obtained by normalizing the length of the major axis of the area by the length of the diagonal.
Short axis length (prop_minaxis)	The value obtained by normalizing the length of the short axis of the area by the length of the diagonal.
Eccentricity (prop_ecc)	Eccentricity of the area.
Area ratio (prop_solidity)	Area ratio of the area.

Combine all the features: 13 + 40 + 6 = 59 features.

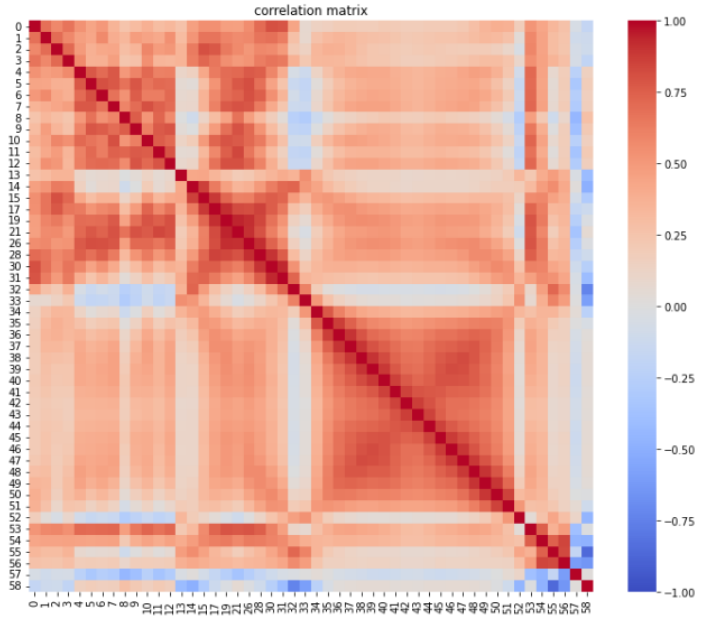
5. Modeling (1/4)

- Handling multi-collinearity
  Correlation matrix



59 features

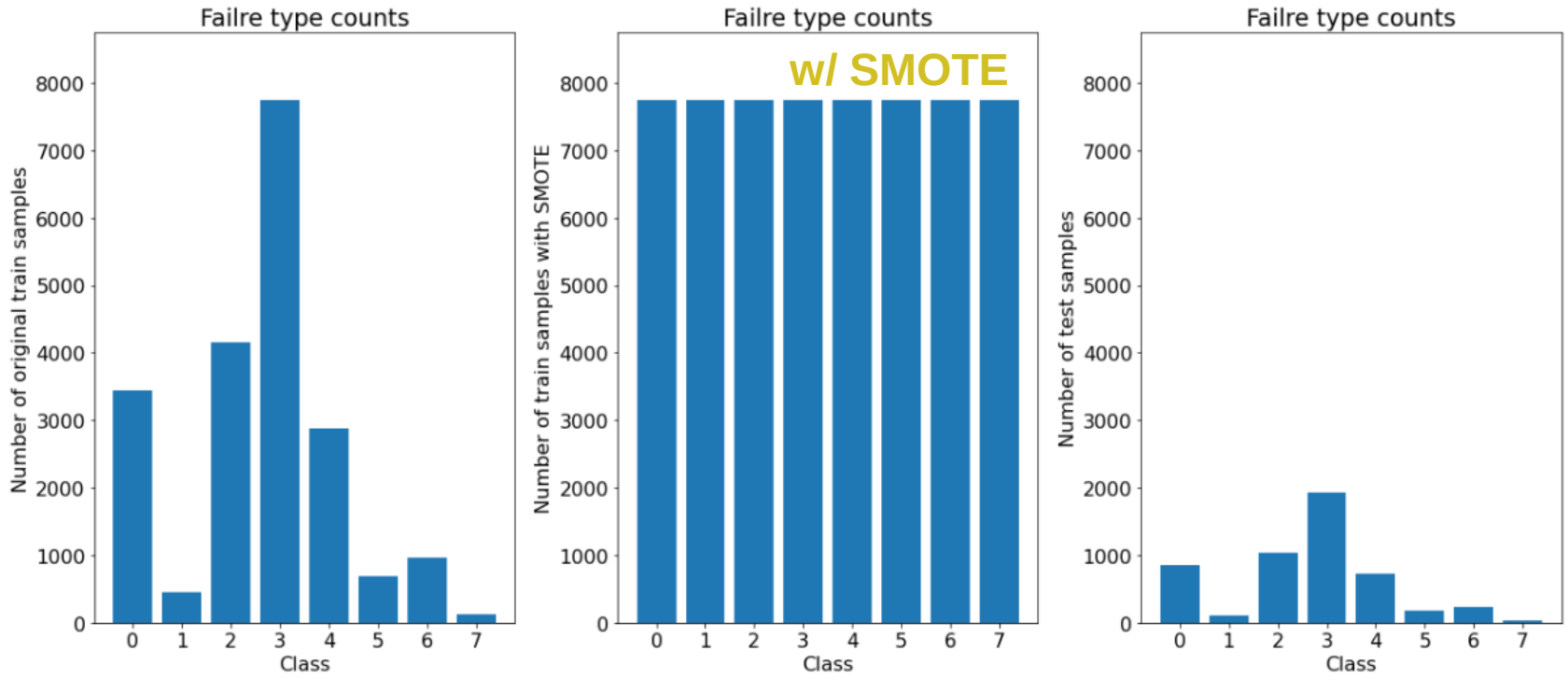
Remove features
VIF > 10



50 features

5. Modeling (2/4)

- Handling data imbalance [5]



Oversampling method SMOTE to eliminate the imbalance data

5. Modeling (3/4)

- Modeling results

  Applied three models: Random Forest, XGBoost, and LightGBM
  GridsearchCV is applied for hyperparameter optimization.

model	training accuracy	test accuracy	processing time
Random Forest	0.99997	0.90263	388.3 sec
XGBoost	0.99997	0.91360	674.8 sec
LightGBM	0.99997	**0.91810**	**65.1 sec**

LightGBM is the highest test accuracy and learning efficiency

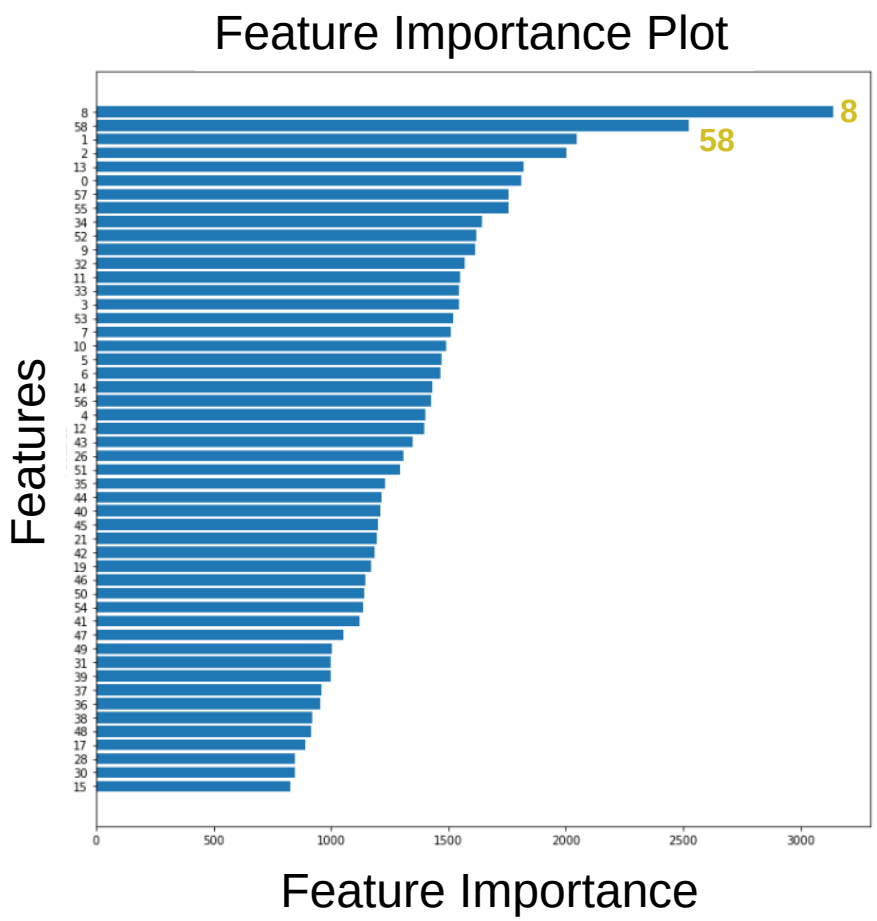
5. Modeling (4/4)

- Feature importance

The feature importance of the best model;

best model: LGBMClassifier(force_row_wise=True,
learning_rate=0.2, n_estimators=200,
num_leaves=63, random_state=42, verbose=-1)

Features 8 and 58 is outstanding.
Feature 8 is the density at the center of the wafer.
Feature 58 is eccentricity.



Quantify the importance of features on defect maps

6. Results and Analysis

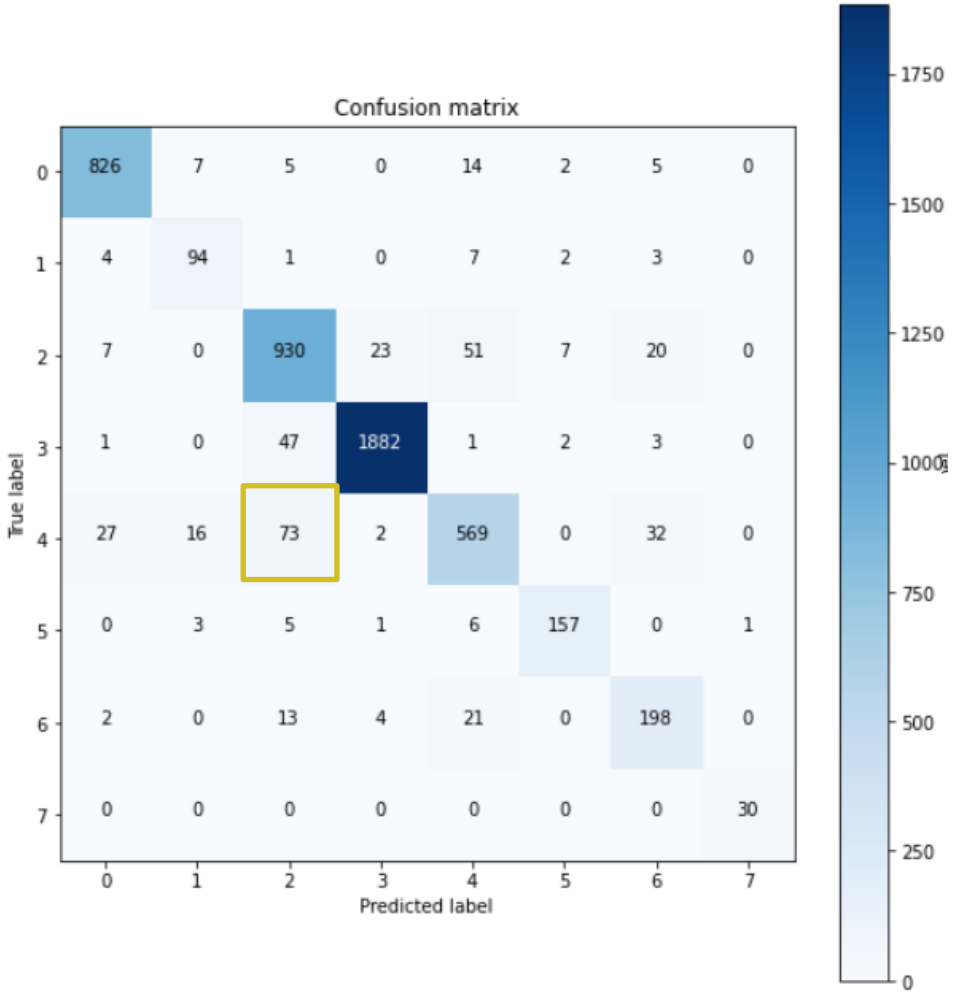
- Confusion matrix

The frequency of predicting {2: 'Donut'} incorrectly as {4: 'Edge-Ring'} is highest.

These two classes have similar tendencies,

it can be expected to be difficult for machine learning.

Overall, accuracy and recall are good results



Confusion matrix shows good prediction performance

7. Discussion and Conclusion

- The project aimed to identify defect characteristics on semiconductor wafers and improve yield technology.
- Random Forest, XGBoost, and LightGBM achieved high accuracy (**LightGBM the highest at 91.81%**) for identifying defective maps.
- Various feature engineering and hyperparameter optimization improved model accuracy.
- Future tasks involve further hyperparameter optimization and applying deep learning like CNNs for accuracy enhancement.

References

[1] M. -J. Wu, J. -S. R. Jang and J. -L. Chen, "Wafer Map Failure Pattern Recognition and Similarity Ranking for Large-Scale Data Sets," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 1, pp. 1-12, Feb. 2015, doi: 10.1109/TSM.2014.2364237.

[2] Mengying Fan, Qin Wang and B. van der Waal, "Wafer defect patterns recognition based on OPTICS and multi-label classification," *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Xi'an, China, 2016, pp. 912-915, doi: 10.1109/IMCEC.2016.7867343.

[3] Dataset, http://mirlab.org/dataSet/public/, or https://www.kaggle.com/datasets/qingyi/wm811k-wafer-map

[4] Wikipedia contributors. (2023, August 15). Radon transform. In Wikipedia, The Free Encyclopedia. Retrieved 01:00, September 2, 2023, from https://en.wikipedia.org/w/index.php?title=Radon_transform&oldid=1170586091

[5] Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. (2002-06-01). "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research*. 16: 321–357. arXiv:1106.1813. doi:10.1613/jair.953. ISSN 1076-9757. S2CID 1554582.

[6] Breiman, Leo (2001). “Random Forests”. Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324.

[7] Brownlee, Jason (March 31, 2020). "Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost".

[8] GitHub repository, https://github.com/DaisakuIkoma/CU_MSDS_SVML.git.


I am a data science Buff.
University of Colorado Boulder
Be Boulder.