Synthetic Data Driven Random Forest Modelling Approach for Predicting Front-End-of-the-Line Silicon Wafer Cleaning Damage

Table of Contents

0. Abstract

Semiconductor manufacturing has become a pivotal industry, with semiconductors increasingly underpinning advancements in medicine, finance, and engineering. Front-end-of-the-line (FEOL) cleaning processes are essential for removing contaminants on the silicon wafer surface, ensuring the longevity and efficiency of the semiconductor device. However, novel semiconductor cleaning processes risk damaging a silicon wafer surface due to unforeseen chemical or mechanical aggressiveness, thermal damage, or surface modification. This study proposes a random forest (RF) regression machine learning (ML) model to predict the probability a FEOL cleaning process will damage a silicon wafer surface. As comprehensive data was inaccessible behind confidential or paywalled research papers, the model was trained off synthetic data, generated and labelled based on publicly available statistics on various indicators of semiconductor health. Despite limited access to real-world data diminishing the accuracy of synthesised data, the fitted RF regressor proved highly accurate (Mean Standard Error = 0.000222, Mean Absolute Error = 0.010073, R^2 Score = 0.995003), with graphical analysis delineating it effective at predicting damage probabilities < 0.2, however tending to slightly underpredict damage probability values > 0.7. The findings indicate that wafer damage probability can be reliably predicted as a function of the FEOL semiconductor cleaning method employed, however access to comprehensive datasets on the effects of different silicon wafer cleaning methods is essential to develop a more accurate model.

1. Literature Review

1.1 *Relevance and Current Understanding*

Advancements in semiconductor manufacturing have become increasingly vital in the development of upcoming technologies and the progression of the global economy. With major investment in semiconductor device production technologies by corporations and governments alike, the industry has ballooned to a current market value of USD 755.28 billion as of 2025 (Fortune Business Insights, 2024).

Particle, organic, and metallic impurities in the silicon wafer during production hinders subsequent manufacturing steps such as lithography and etching, consequently negatively affecting the final performance of the chip (Stanford Advanced Materials, 2025). Wafer purity impacts the yield of products (Wafer World, February 12 2024), as uniformity in the crystalline structure allows for maximum yield fit for consumer sale, as well as the overall longevity and stability of a semiconductor device (MSR-FSR, n.d.).

Front-end-of-the-line (FEOL) cleaning describes cleaning processes preceding critical high-temperature processes, preventing contaminants from diffusing into silicon (Wagner, 2024). Multitudes of FEOL cleaning methods persist to mitigate the adverse effects of these impurities (Stanford Advanced Materials, 2019). However novel cleaning processes risk damaging thin films and microstructures on the silicon wafer surface itself (DRex, 2024), thereby hindering the overall performance of the semiconductor (JAS, 2025) (Joyce et. al, 2014).

There is currently no quantifiable way to predict how a new FEOL silicon wafer cleaning method will impact the likelihood of damage on the final semiconductor device. This study proposes a RF regressor to accurately predict probability a FEOL cleaning process will damage a silicon wafer surface.

1.2 *Criterion Decisions and the use of Synthetic Data*

‘Synthetic data’ represents artificially generated data based on limited collected real world data. When data are scarce, unavailable, or of poor quality, synthetic data can be used, for example, to improve the performance of machine learning (ML) models (Figueira and Vaz, 2022).

Much of all comprehensive available data on the effects of silicon wafer cleaning methods on the likelihood of final product wafer damage was inaccessible due to commercial or governmental reasons, or only accessible to paying subscribers (so-called ‘paywalled’) . Consequently, the data in this study was synthetically generated to enable the training of an accurate ML model despite logistical limitations.

The methodology outlines a detailed description of how the synthetic data was generated and labelled. Additionally, the model has been designed to be easily fitted to and trained on real world data once acquired.

2. Scientific Research Question

Based on the type of FEOL cleaning method used, can the likelihood of a given silicon wafer being damaged be quantitatively predicted?

3. Scientific Hypothesis

The damage probability of a given silicon wafer based on the characteristics of the FEOL cleaning process used in its production can be predicted based on a distinct set of criteria and adequate data on the effects of different semiconductor cleaning processes on a set silicon wafer process.

4. Methodology

The code history for this investigation can be found at: https://github.com/Art4v/science_extension

4.1 *Synthetic Data*

Semiconductor manufacturing is an immensely competitive field (University of South Florida, 2023). Consequently, comprehensive data published on the effects of modern semiconductor cleaning methods is often paywalled or privately published.

Data fitted to this predictive model has been synthetically generated. Synthetic data sets emulate certain key information found in the actual data and provide the ability to draw valid statistical inferences, allowing widespread access to data for analysis while mitigating confidentiality concerns (E. Raghunathan, 2021). Synthetic data has been used extensively in past disciplines such as healthcare (Gonzales et al., 2023), computer vision (Zewe, 2022), and finance (K. Potluru et al., 2024).

It must be acknowledged that, however, neither the reliability nor accuracy of synthetic data used can be guaranteed (Gartner, 2024) nor that the criteria considered is a comprehensive list of all factors that affect the likelihood of damage of a given semiconductor. However, the synthetic data enables training of this ML model despite a lack of publicly available data.

4.1.1 *Generation*

The synthetic data generated for this experiment is based on limited results available from publicly available research papers and commercial publications on the effects of cleaning methods on silicon wafers. The list of criteria considered is not comprehensive, however, serves to emulate potential criteria that would have needed to be collected for an accurate model. (See Appendix A)

Means and standard deviations for criteria assessing different industrial cleaning processes were derived from publicly available experimental data, with slight adjustments to simulate variations across different methods. (See Appendix B)

Synthetic samples were generated based on this criteria. For a given synthetic sample, a random cleaning method was selected, and criterion values were generated based on a normal distribution of its mean and standard deviation. 50 000 synthetic samples were generated and saved to a set random state to ensure reproducibility of data.

4.1.2 *Labelling*

To label the data with an appropriate damage probability, a logistic regression used weightings for each criterion based on publicly available industrial thresholds to emulate expert conclusions on criterion significance on wafer damage probability. The criterion weighting derived and their respective justifications can be found in Appendix C.

Based on this weighting, the logistic regression model calculated a damage probability for a given semiconductor cleaning method.

4.1.4 *Random Noise*

To simulate random and human error in results found in experimental results, the generated criterion (excluding the damage probability) was altered with 40% random noise.

4.2 *Model Training*

A Random Forest (RF) regressor (see 4.2.2) was trained on the synthetic data to output a probability score estimating the likelihood of damage caused by a semiconductor cleaning method to a silicon wafer. RF was selected due to its capacity to capture non-linear patterns and trends (Verdiguier et al., 2020), handle imbalanced data and variables with missing values, and reduce the risk of overfitting (Salman et al., 2024), making it well-suited to training on real-world data once available.

4.2.1 *Train-Test Split*

The synthetic data was randomly separated by an 80:20 ratio into training and testing data respectively, as according to industry standards (Ahmed, 2022). Training data was initially fed into the model to allow it to identify patterns and non-linear relationships between criterions and damage probability. Testing data evaluated the accuracy of the model by comparing its predictions based on input criterion to recorded observations.

4.2.2 *Random Forest Regression*

RF is a special type of model ensembling algorithm. Ensemble learning is a ML paradigm that enhances the accuracy of predictions by a model by combining the outputs of multiple models (Özen, 2024). RF uses a bootstrap aggregating (bagging) method of model ensembling, which trains multiple models on different bootstrap samples of the training data and aggregates their predictions (Kwak et al., 2022).

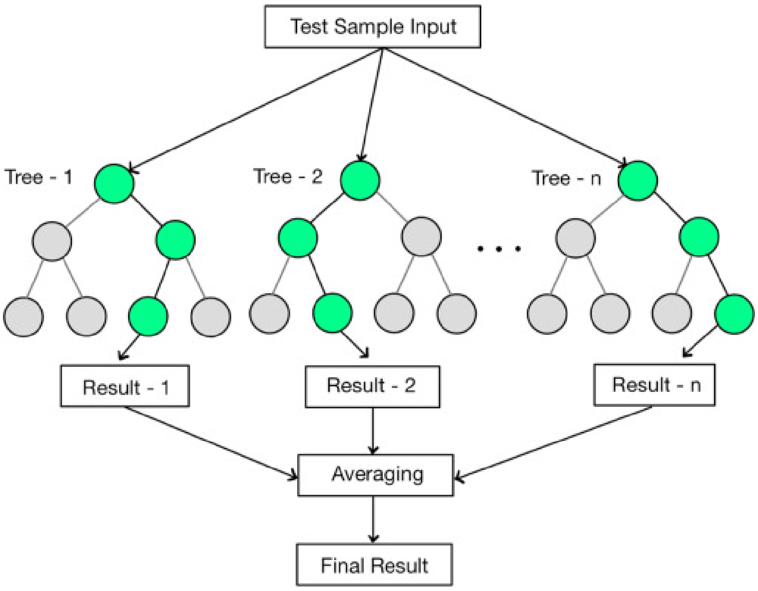

Figure 1. *Visualization of a RF regression model. Image courtesy of Kwak et al., 2022.*

RF ensembles a set number of learner decision trees. Decision trees depict a tree-like structure wherein the internal nodes represent the features, while leaf nodes comprise the class labels or predicted values. A prediction is made by traversing the path of if-else conditions from the root to one of the leaf nodes depending on the feature values of the input (Awad and Fraihat, 2023).

RF regression involves averaging the final numerical results of all the decision trees in the model.

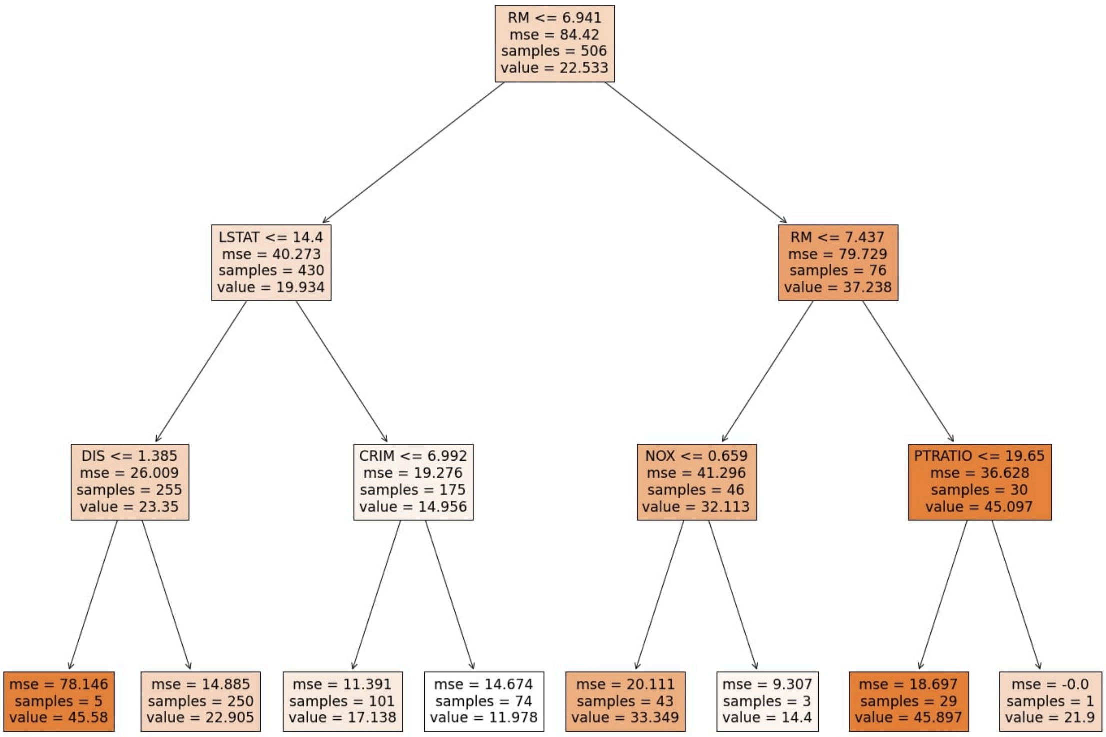

Figure 2. *Visualization of an individual decision tree in the RF regressor. Image courtesy of Płoński, 2020.*

In this RF regression model, each decision tree had a random number of decision nodes and magnitude of branching to increase model diversity and strengthen model performance. 30 decision trees were created and trained based on 40 000 samples of randomly selected training data with replacement. The individual decision trees were not correlated to minimise variance and prevent overfitting in the ensemble RF regressor model (Ganesh N. et al., 2021). The mean of all the 30 ensembled learner decision trees was the output prediction of the RF regressor.

4.3 *Model Testing*

The model was evaluated on a test set split from the synthetic dataset. Key metrics — Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 — were calculated. MSE measures the average squared difference between predicted and observed values, MAE measures the average magnitude of these differences, and R^2 indicates the proportion of variability in the target variable explained by the model.

Additionally, visualisations of predicted vs actual values, residuals, q-q plots, and feature importances were generated and collected for further insight into model performance (see discussion).

5. Results

5.1 *Numerical Statistics*

Mean Squared Error (MSE)	0.000222
Mean Absolute Error (MAE)	0.010073
R^2 Score (R2)	0.995003

Table 8. *RF Regression Model Performance Evaluation Metrics for Random State = 0.*

5.2 *Graphical Analysis*

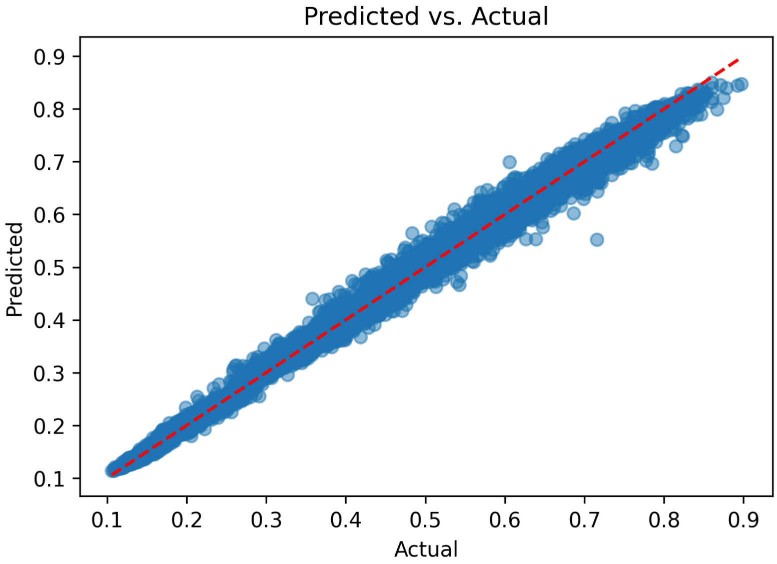

Figure 3. *Scatter plot showing relationship between predicted vs actual values for the damage probability score of a given FOEL semiconductor cleaning method. The red dotted line serves as a reference for perfect linearity.*

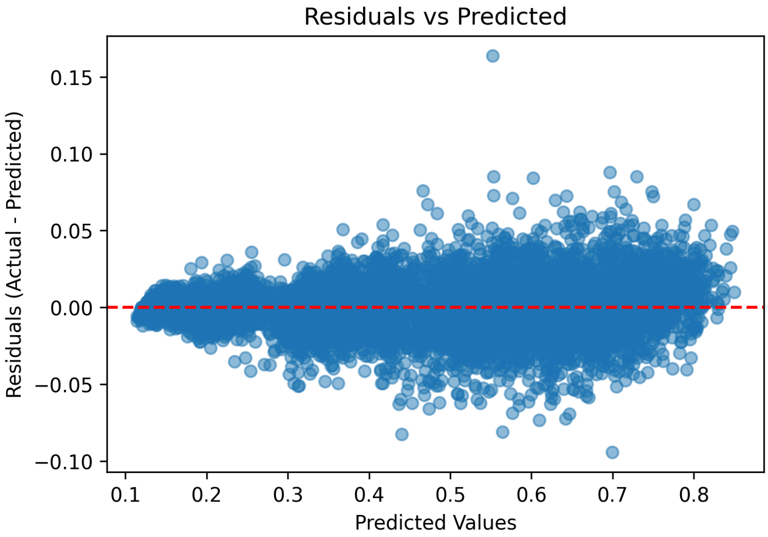

Figure 4. *Residuals plot showing relationship between predicted values for damage probability and their deviation from the actual reported value of a given semiconductor cleaning method.*

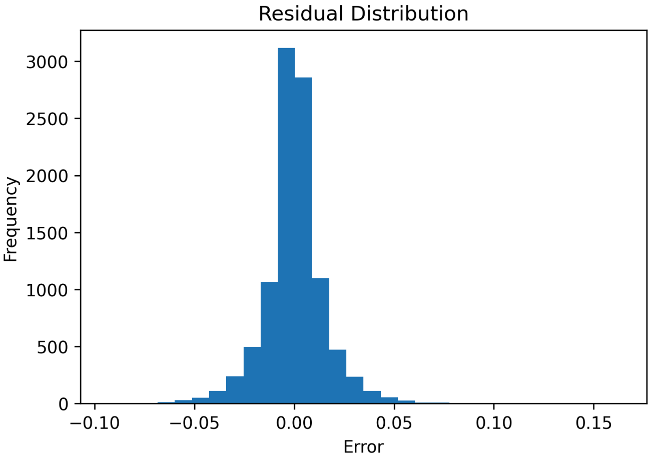

Figure 5. *Residuals distribution showcasing the frequency of different magnitudes of deviations between predicted and reported values for damage probability of a given semiconductor cleaning method.*

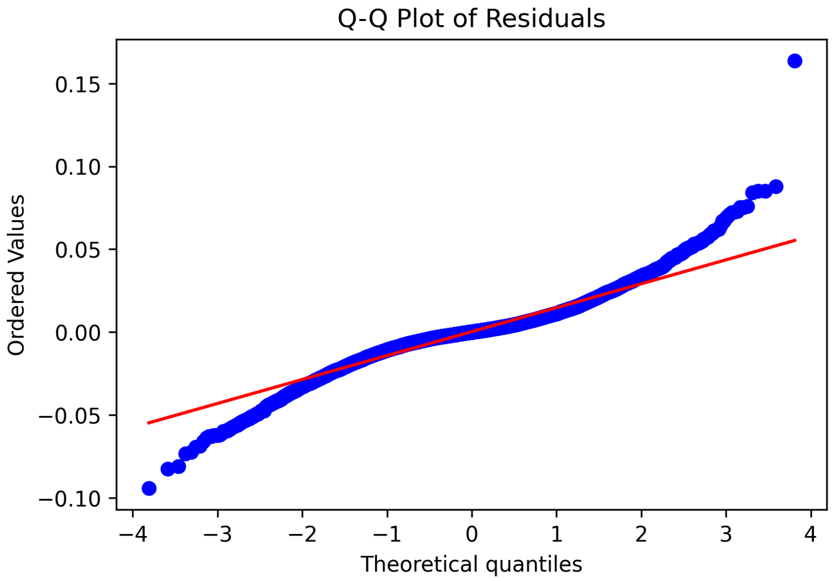

Figure 6. *Q-Q plot showing the relationship between theoretical quantiles and ordered values of the residuals of predicted versus actual damage probabilities. The red line serves as a reference line for a perfect normal distribution of datapoints.*

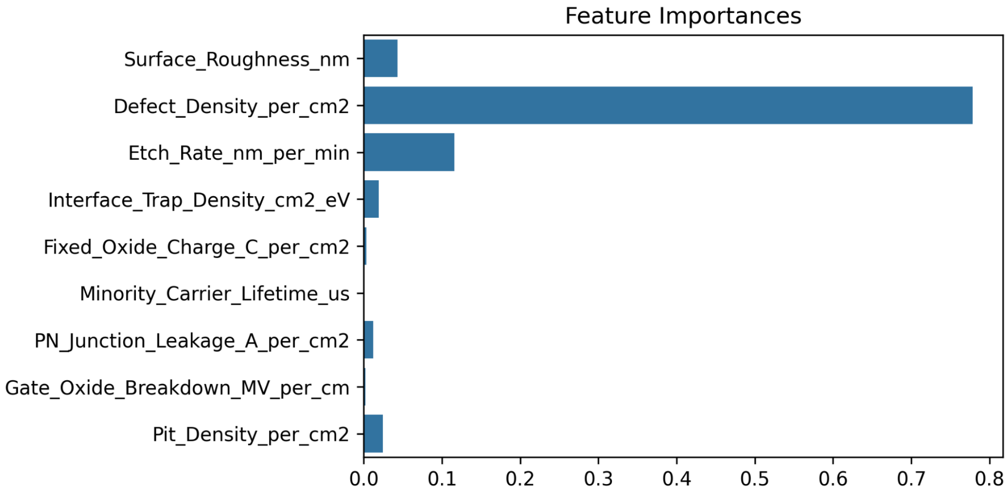

Figure 7. *Feature Importance bar graph showing the interpreted weighting of the model for how different criteria affect the damage probability for a given semiconductor cleaning method.*

6. Discussion

6.1 *Derivation of Synthetic Data*

6.1.1 *Generation*

Without real-world data on the effects of different semiconductor cleaning methods, the data criterion generated cannot be a comprehensive list of all the indicators of damage on a silicon wafer. Furthermore, restricted access to reliable sources to base the means and standard deviations for different FEOL silicon wafer cleaning processes further detriments the reliability and accuracy of the data generated.

However, the generation of synthetic data creates a high quality dataset. The dataset does not contain missing values or outliers, provides equal representation for different class labels, and is large in scale (= 50 000 samples), enabling the resultant trained predictive model to be highly accurate and reliable.

6.1.2 *Labelling*

The labelling of synthetic data, similarly to its generation, is hindered in reliability and accuracy due to limited access to reliable sources to base weighting for different FEOL silicon wafer cleaning processes. The experiment would benefit from data labelling conducted by industry experts, facilitating a much more accurate labelling of the synthetic data.

However, the use of logistic regression weighted according to researched industrial standards, alongside random noise to emulate real-world experimental error, allowed for relatively accurate labelling given the limitations of the experiment.

6.2 *Evaluation and Analysis of Model Performance*

6.2.1 *Numerical Evaluation*

From Table 8, The MSE = 0.000222 is significantly lower than the industrial standard of 0.05, and the MAE = 0.010073 approximates to 0.1, indicating minimal differences between predicted and actual values and a generally high accuracy. The R2 value of this model = 0.995003 indicates high explanatory power of predictions with minimal noise, however suggests that the model may be overfitted.

Overall, numerical evaluation showcases promising predictive accuracy, however further graphical analysis will enable insight into the strengths and weaknesses of the ML model.

6.2.2 *Predicted vs Actual Scatterplot*

The predicted vs actual scatterplot (Figure 3) shows the trend of predicted value points generally lying close to the perfect prediction line. For damage probability values < 0.2, the predicted value is closest to the actual value with little to no scatter, indicating high predictive accuracy. For damage probability values 0.2 < 0.7, there is a substantial increase in the scatter of predictive values and there several outlier predictions both over and under the perfect prediction line, however the model neither consistently underpredicts or overpredicts for these ranges, indicating a lower but still high predictive accuracy for this range. However, for damage probability values > 0.7, we notice data points trend below the perfect prediction line, indicating consistent underprediction of values compared to actuality.

Overall, the model has a high predictive capacity, maintaining predictions close to the perfect prediction line. However, the model has high heteroscedasticity, and its accuracy in predictions dwindles as the likelihood of damage increases, as the model slowly begins to underpredict.

6.2.3 *Residuals Plot*

An ideal residuals plot shows residuals to be randomly distributed across all predicted values. Looking at Figure 4, the residuals show minimal spread towards the lower predicted values < 0.3, however the magnitude of spread slightly increases past this point. We see several outliers for predicted values > 0.4, and the breath of the residuals plot increases past this point, indicating slight heteroscedasticity across predictions made by the model.

Overall, the residuals plot of predictions shows magnitude of residuals increase with increasing damage probability, however the predicted values generally show high homoscedasticity.

6.2.4 *Residuals Distribution*

The residuals distribution enables the interpretation of the distribution shape of prediction errors. In Figure 5, the residuals histogram for the predictive model is mostly normally distributed and monomodal, with a very slight skew to the left and small tails. This indicates that the errors the model makes are mostly random and unbiased, with a very slight tendency to underestimate damage probabilities and make extreme errors.

6.2.5 *Q-Q Plot of Residuals*

The Q-Q plot of residuals (Figure 6) is a visual representation of the distribution of the residuals of the model against a theoretical normal distribution (indicated by the red line). For the model, theoretical quintiles > -2 and < 2 closely follow the red line, indicating that residuals made for predictions within this range closely follow a normal distribution. However, the overall inverted ‘S’ shape of the curve depicts the theoretical quantiles < -2 significantly below and > 2 significantly above the red line. This indicates that the model has fewer extreme negative values and more extreme positives than a normal distribution would predict. This evidence further supports that the model is more likely to underpredict probabilities.

6.2.6 *Feature Importance Graph*

The feature importance plot showcases the deduced weighting of the model for each of the individual factors influencing its decisions (Figure 7). Examining these results compared to the initial weightings indicates while generating synthetic data (Table 7, in Appendix C), we see that the model abnormally considered “Defect Density” as a significant indicator for a damage probability, while giving little consideration to “Fixed Oxide Charge ”, “Minority Carrier Lifetime”, and “Gate Oxide Breakdown”. This lies in stark contrast to the assigned weighting of the synthetic data favouring “Surface Roughness” as having the highest feature importance.

However, it should be noted that the criteria being judged upon is arbitrary, and no conclusions can be drawn from this graphical analysis until the model is trained with real world experimental data.

7. Conclusion

This study explored the use of ML techniques to estimate the probability of damage done to silicon wafers based on the FEOL cleaning processes used in its manufacturing. Unable to access adequate real world data on the characteristics of different cleaning processes, the study generated and labelled synthetic data based on a python script to mimic means and standard deviations for set criterion, while adding noise to emulate experimental error in the dataset. The study trained a RF regression model with 30 estimators based on this synthetic data. The model was found to have a high accuracy, generally higher for lower damage probabilities < 0.2 , but had a slight tendency to underestimate its predictions for predictions > 0.7. Once granted access to real world experimental data, the model can be easily fitted to make accurate predictions on damage probability based on the type of cleaning process used. Thus, this provides evidence that it is feasible to predict the likelihood a given silicon wafer will be damaged based on the type of FEOL semiconductor cleaning method used.

8. References

Ahmed (2025). The Motivation for Train-Test Split. [online] https://medium.com. Available at: https://medium.com/@nahmed3536/the-motivation-for-train-test-split-2b1837f596c3.

Awad, M. and Fraihat, S. (2023). Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. Journal of Sensor and Actuator Networks, [online] 12(5), p.67. doi:https://doi.org/10.3390/jsan12050067.

Best Cleaning (2012). The evolution of silicon wafer cleaning technology. [online] J. Electrochem. Soc. Available at: https://www.academia.edu/1877592/The_evolution_of_silicon_wafer_cleaning_technology.

Celler, G. (1999). Etching of Silicon by the RCA Standard Clean 1. [online] Available at: https://www.researchgate.net/publication/244669400_Etching_of_Silicon_by_the_RCA_Standard_Clean_1.

Chanana, R., Dwivedi, R. and Srivastava, S. (2009). SILICON-WAFER CLEANING WITH CF4/H2 PLASMA AND ITS EFFECT ON THE PROPERTIES OF DRY THERMALLY GROWN OXIDE. [online] Epa.gov. Available at: https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/6587410.

D. Martin Knotter, Gendt, S. de, Mertens, P.W. and Heyns, M.M. (2000). Silicon Surface Roughening Mechanisms in Ammonia Hydrogen Peroxide Mixtures. Journal of The Electrochemical Society, [online] 147(2), p.736. doi:https://doi.org/10.1149/1.1393261.

DRex (2024). Semiconductor Cleaning Processes: Methods and Importance - DRex Electronics. [online] DRex Electronics. Available at: https://www.icdrex.com/semiconductor-cleaning-processes-methods-and-importance.

Du, C., Zhao, Y. and Li, Y. (2023). Effect of Surface Cleaning Process on the Wafer Bonding of Silicon and Pyrex Glass. Journal of Inorganic and Organometallic Polymers and Materials, 33(3), pp.673–679. doi:https://doi.org/10.1007/s10904-022-02510-x.

Eurika (2025). How surface roughness affects lithography resolution and yield. [online] Patsnap.com. Available at: https://eureka.patsnap.com/article/how-surface-roughness-affects-lithography-resolution-and-yield.

Figen Özen (2024). Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey. Heliyon, 10(4), pp.e25746–e25746. doi:https://doi.org/10.1016/j.heliyon.2024.e25746.

Figueira, A. and Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. Mathematics, 10(15), p.2733. doi:https://doi.org/10.3390/math10152733.

Fortune Business Insights (2022). Semiconductor Market Size & Share | Industry Growth [2020-2027]. [online] www.fortunebusinessinsights.com. Available at: https://www.fortunebusinessinsights.com/semiconductor-market-102365.

Gale, G. (1996). Defect reduction and cost savings through re-inventing RCA cleans. [online] IEEE/SEMI Advanced Semiconductor Manufacturing Conference. Available at: https://www.academia.edu/62709902/Defect_reduction_and_cost_savings_through_re_inventing_RCA_cleans.

Garnier, P., Galbes, H. and Laurent Viravaux (2024). Surface Damage by Physical Cleans during Semiconductors Manufacturing. ECS Transactions, 114(1), pp.73–81. doi:https://doi.org/10.1149/11401.0073ecst.

Gartner (2023). Synthetic Data. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/synthetic-data.

Gonzales, A., Guruswamy, G. and Smith, S.R. (2023). Synthetic data in health care: A narrative review. PLOS Digital Health, 2(1), p.e0000082. doi:https://doi.org/10.1371/journal.pdig.0000082.

Han, Z., Keswani, M. and Raghavan, S. (2013). Megasonic Cleaning of Blanket and Patterned Samples in Carbonated Ammonia Solutions for Enhanced Particle Removal and Reduced Feature Damage. IEEE Transactions on Semiconductor Manufacturing, [online] 26(3), pp.400–405. doi:https://doi.org/10.1109/TSM.2013.2271641.

Waferpro (2024). How Silicon Wafer Defects Impact Device Performance | WaferPro. [online] WaferPro. Available at: https://waferpro.com/how-silicon-wafer-defects-impact-device-performance.

Cady, W.A. and Varadarajan, M. (2025). Radware Bot Manager Captcha. [online] Available at: https://iopscience.iop.org/article/10.1149/1.1836950/meta.

Izquierdo-Verdiguier, E. and Zurita-Milla, R. (2020). An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. International Journal of Applied Earth Observation and Geoinformation, 88, p.102051. doi:https://doi.org/10.1016/j.jag.2020.102051.

JAS (2025). JAS Precision Electronics. [online] JAS Precision Electronics. Available at: https://www.jas8.com.tw/news/understanding-and-removing-particles-on-silicon-wafers.

Joyce, R., Singh, K., Varghese, S. and Akhtar, J. (2015). Effective cleaning process and its influence on surface roughness in anodic bonding for semiconductor device packaging. Materials Science in Semiconductor Processing, 31, pp.84–93. doi:https://doi.org/10.1016/j.mssp.2014.11.002.

Kim, K. (2012). Reliable CMOS VLSI Design Considering Gate Oxide Breakdown. [online] Available at: https://www.thinkmind.org/articles/cenics_2012_2_50_70062.pdf.

Kwak, S., Kim, J., Ding, H., Xu, X., Chen, R., Guo, J. and Fu, H. (2022). Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model. Journal of Materials Research and Technology, [online] 18, pp.520–530. doi:https://doi.org/10.1016/j.jmrt.2022.02.108.

Meuris, M., Arnauts, S., Cornelissen, I., Kenis, K., Lux, M., Gendt, S.D., Mertens, P.W., I. Teerlinck, Vos, R., Loewenstein, L., Heyns, M.M. and Wolke, K. (2023). Implementation of the IMEC-cleaning in advanced CMOS manufacturing. CRC Press eBooks, pp.57–67. doi:https://doi.org/10.1201/9780429070716-6.

Min, T., Park, B., Kang, S., Gweon, G., Kim, Y. and Yeom, G. (2009). Improvement of surface roughness in silicon-on-insulator wafer fabrication using a neutral beam etching. [online] Available at: https://swb.skku.edu/_res/pnpl/etc/2009-10.pdf?utm_source.

MKS (n.d.). Wafer Surface Cleaning. [online] www.mks.com. Available at: https://www.mks.com/n/wafer-surface-cleaning.

MSR-FSR (2024). What is the cleaning process of semiconductors?. [online] MSR-FSR. Available at: https://msr-fsr.com/what-is-the-cleaning-process-of-semiconductors.

N., G., Jain, P., Choudhury, A., Dutta, P., Kalita, K. and Barsocchi, P. (2021). Random Forest Regression-Based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes. Processes, 9(11), p.2095. doi:https://doi.org/10.3390/pr9112095.

Nadun Sinhabahu, Li, K.S.-M., Li, J.-D., Wang, J.R. and Wang, S.-J. (2022). Yield-Enhanced Probe Head Cleaning with AI-Driven Image and Signal Integrity Pattern Recognition for Wafer Test. [online] doi:https://doi.org/10.1109/itc50671.2022.00071.

Płoński, P. (2020). Visualize a Decision Tree in 5 Ways with Scikit-Learn and Python. [online] MLJAR. Available at: https://mljar.com/blog/visualize-decision-tree.

Potluru, V.K., Borrajo, D., Coletta, A., Dalmasso, N., El-Laham, Y., Fons, E., Ghassemi, M., Gopalakrishnan, S., Gosai, V., Kreačić, E., Mani, G., Obitayo, S., Paramanand, D., Raman, N., Solonin, M., Sood, S., Vyetrenko, S., Zhu, H., Veloso, M. and Balch, T. (2024). Synthetic Data Applications in Finance. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2401.00081.

Raghunathan, T.E. (2020). Synthetic Data. Annual Review of Statistics and Its Application, 8(1). doi:https://doi.org/10.1146/annurev-statistics-040720-031848.

Reinhardt, K.A. and Kern, W. (2018). Handbook of silicon wafer cleaning technology. Kidlington William Andrew Applied Science Publishers, Imprint Of Elsevier.Sahoo, B.N., Han, S.Y., Kim, H.-T., Ando, K., Kim, T.-G., Kang, B.-K., Klipp, A., Yerriboina, N.P. and Park, J.-G. (2021). Chemically controlled megasonic cleaning of patterned structures using solutions with dissolved gas and surfactant. Ultrasonics Sonochemistry, [online] 82, p.105859. doi:https://doi.org/10.1016/j.ultsonch.2021.105859.

Salman , H.A., Kalakech , A. and Steiti, A. (2024). View of Random Forest Algorithm Overview. [online] Mesopotamian.press. Available at: https://journals.mesopotamian.press/index.php/BJML/article/view/417/289.

Sato, Y., Shibata, S., Nishimura, K., Yamasaki, M., Murakami, M., Urabe, K. and Koji Eriguchi (2022). Predicting the effects of plasma-induced damage on p–n junction leakage and its application in the characterization of defect distribution. Journal of Vacuum Science & Technology B Nanotechnology and Microelectronics Materials Processing Measurement and Phenomena, [online] 40(6). doi:https://doi.org/10.1116/6.0002181.

Sopori, B., Devayajanam, S. and Basnyat, P. (2015). Using minority carrier lifetime measurement to determine saw damage characteristics on Si wafer surfaces. 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), [online] pp.1–6. doi:https://doi.org/10.1109/pvsc.2015.7355744.

Sputter, S. (2019). Silicon Wafer: 4 Types of Wet Cleaning Method - SAM Sputter Targets. [online] SAM Sputter Targets. Available at: http://www.sputtering-targets.net/blog/4-types-of-wet-cleaning-method-of-silicon-wafer/.

Stanford Advanced Materials (2025). The Impact of Silicon Wafer Quality on Semiconductor Performance and Reliability. [online] www.samaterials.com. Available at: https://www.samaterials.com/the-impact-of-silicon-wafer-quality-on-semiconductor-performance-and-reliability.html.

Tomohisa Tokura (2024). Innovative Techniques for Organic Residue Removal in Semiconductor Manufacturing Using Backside Brush Scrubber Clean. [online] doi:https://doi.org/10.13140/RG.2.2.19350.97606.

Tseng, A. (2006). Effects of surface roughness and oxide layer on wafer bonding strength using transmission laser bonding technique. [online] Available at: https://www.researchgate.net/publication/224636038_Effects_of_surface_roughness_and_oxide_layer_on_wafer_bonding_strength_using_transmission_laser_bonding_technique.

University of South Florida (2023). What is semiconductor technology and why is it important? [online] Usf.edu. Available at: https://www.usf.edu/continuing-education/lifelong-learning/news/2023/what-is-semiconductor-technology-and-why-is-it-important.aspx.

Wafer World (2021). Silicon Wafer Cleaning: Methods and Techniques. [online] Waferworld.com. Available at: https://www.waferworld.com/post/how-are-silicon-wafers-cleaned.

Wagner, D. (2024). Modutek Corporation. [online] Modutek. Available at: https://www.modutek.com/silicon-wafer-cleaning-removing-contaminants-to-improve-yields.

Yu, S., White, M.H. and Agarwal, A.K. (2021). Experimental Determination of Interface Trap Density and Fixed Positive Oxide Charge in Commercial 4H-SiC Power MOSFETs. IEEE Access, 9, pp.149118–149124. doi:https://doi.org/10.1109/access.2021.3124706.

Zewe, A. (2022). In machine learning, synthetic data can offer real performance improvements. [online] MIT News | Massachusetts Institute of Technology. Available at: https://news.mit.edu/2022/synthetic-data-ai-improvements-1103.

9. Appendices

Appendix A:

Criterion	Summary and Source
Surface Roughness (nm RMS)	The root-mean-square variation of surface height on the wafer (nm) (D. Martin Knotter et al., February 2000).
Defect Density (#/cm^2)	The number of surface defects per unit area (#/cm^2). (Zhenxing Han et al., August 2013).
Etch Rate (nm/min)	The rate of silicon removal by the cleaning solution (Å/min or nm/min) (George Celler, January 1999).
Interface Trap Density (cm^{-2}.eV^{-1})	The density of electronic trap states at the Si–SiO$_2$ interface (cm^{-2}·eV^{-1}) (RK Chanana et al., 1992).
Fixed Oxide Charge (C/cm^2)	Net fixed charge in the gate oxide per area (C/cm^2) (RK Chanana et al., 1992).
Minority Carrier Lifetime (µs)	The average time minority carriers survive before recombining (µs) (MKS, n.d.)
Leakage Current (A/cm^2)	Reverse-bias leakage current of a p–n junction (A/cm^2). (MKS, n.d.)
Breakdown Field (MV/cm)	The electric field at which gate oxide fails (MV/cm). (D. Martin Knotter et al., February 2000)
Pit Density (pits/cm^2)	The number of microscopic etch pits per area (pits/cm^2). (D. Martin Knotter et al., February 2000)

Table 1. *Criteria considered when generating synthetic dataset*

Appendix B:

Method: SC-1

Criterion	Mean	Standard Deviation
Surface Roughness (nm RMS)	1.0	0.2
Defect Density (#/cm^2)	5e5	1e5
Etch Rate (nm/min)	10	3
Interface Trap Density (cm^{-2}.eV^{-1})	5e11	1e11
Fixed Oxide Charge (C/cm^2)	5e-6	1e-6
Minority Carrier Lifetime (µs)	20	5
Leakage Current (A/cm^2)	1e-8	5e-9
Breakdown Field (MV/cm)	9	1
Pit Density (pits/cm^2)	500	200

Table 2. *Set means and standard deviations of synthetic data for SC-1*

Method: RCA-1

Criterion	Mean	Standard Deviation
Surface Roughness (nm RMS)	0.8	0.15
Defect Density (#/cm^2)	3e5	8e4
Etch Rate (nm/min)	8	2
Interface Trap Density (cm^{-2}.eV^{-1})	3e11	8e10
Fixed Oxide Charge (C/cm^2)	4e-6	8e-7
Minority Carrier Lifetime (µs)	30	8
Leakage Current (A/cm^2)	5e-9	2e-9
Breakdown Field (MV/cm)	10	0.8
Pit Density (pits/cm^2)	300	150

Table 3. *Set means and standard deviations of synthetic data for RCA-1*

Method: UV/Ozone

Criterion	Mean	Standard Deviation
Surface Roughness (nm RMS)	0.3	0.1
Defect Density (#/cm^2)	1e5	5e4
Etch Rate (nm/min)	0.5	0.2
Interface Trap Density (cm^{-2}.eV^{-1})	1e11	5e10
Fixed Oxide Charge (C/cm^2)	2e-6	5e-7
Minority Carrier Lifetime (µs)	50	10
Leakage Current (A/cm^2)	1e-10	5e-11
Breakdown Field (MV/cm)	12	0.5
Pit Density (pits/cm^2)	50	30

Table 4. *Set means and standard deviations of synthetic data for UV/Ozone*

Method: HF Dip

Criterion	Mean	Standard Deviation
Surface Roughness (nm RMS)	0.5	0.12
Defect Density (#/cm^2)	2e5	6e4
Etch Rate (nm/min)	2	0.5
Interface Trap Density (cm^{-2}.eV^{-1})	2e11	6e10
Fixed Oxide Charge (C/cm^2)	3e-6	6e-7
Minority Carrier Lifetime (µs)	40	7
Leakage Current (A/cm^2)	1e-9	4e-10
Breakdown Field (MV/cm)	11	0.7
Pit Density (pits/cm^2)	150	80

Table 5. *Set means and standard deviations of synthetic data for HF Dip*

<u>*Method: Piranha*</u>

Criterion	Mean	Standard Deviation
Surface Roughness (nm RMS)	1.2	0.25
Defect Density (#/cm^2)	7e5	1.2e5
Etch Rate (nm/min)	12	4
Interface Trap Density (cm^{-2}.eV^{-1})	6e11	1.2e11
Fixed Oxide Charge (C/cm^2)	6e-6	1.2e-6
Minority Carrier Lifetime (µs)	15	4
Leakage Current (A/cm^2)	2e-8	6e-9
Breakdown Field (MV/cm)	8	1.2
Pit Density (pits/cm^2)	600	250

Table 6. *Set means and standard deviations of synthetic data for Piranha*

Appendix C:

Criterion	Weighting	Justification
Surface Roughness (nm RMS)	0.3	Rough surfaces are known to cause significant yield loss and particulate issues in FEOL processing (Eurika, 2025).
Defect Density (#/cm^2)	0.2	Induced defects directly reduce viable die yield (WaferPro, 2024).
Etch Rate (nm/min)	0.1	Excessive etch indicates overdosing of chemicals or process drift (T H Min et al., 2009).
Interface Trap Density (cm^{-2}.eV^{-1})	0.15	Elevated interface trap density degrades electrical performance and reliability (Susanna Yu et al., 2021).
Fixed Oxide Charge (C/cm^2)	0.1	High charge disrupts threshold voltage and electrical stability (Susanna Yu et al., 2021).
Minority Carrier Lifetime (µs)	-0.05	Higher lifetimes imply minimal subsurface damage (NREL, 2015).
Leakage Current (A/cm^2)	0.1	Elevated leakage signals traps or surface damage in the junction zone (Yoshihiro Sato et al., 2022).
Breakdown Field (MV/cm)	0.1	Lower breakdown fields point to compromised gate dielectric integrity (Kyung Ki Kim,

		2012).
Pit Density (pits/cm^2)	0.2	Visible etch pits often indicate subsurface damage, increasing scrap rates (Ampere Tseng, 2006).

Table 7. *Set weightings and justifications for logistic regression labelling*