Using Machine Learning Solutions to Accurately Classify Imbalanced LCM Aging Data to Reduce Defect Rates

Jing Ba, Xiaofeng Wei**, Song Gao*, Shun Lin**, Zhiwei Tan*, Yangxing Liu***

***TCL China Star Optoelectronics Technology Co., Ltd., Shenzhen, China**

****Wuhan TCL Research Co., Ltd., Wuhan, China**

*Jing Ba and Xiaofeng Wei are co-first authors and contributed equally to this work

Abstract

In semiconductor display manufacturing, liquid crystal display module (LCM) aging is a critical process for assessing the longevity, reliability, and performance degradation of devices over time. However, LCM aging data from real production lines are severely imbalanced, which reduces the accuracy of classifying minority classes, causing significant issues at product quality control. This paper proposes a hybrid method that combines machine learning algorithms, cost-sensitive learning, and SMOTE-ENN techniques to address the imbalanced classification problem in the LCM aging process. The method demonstrates high predictive accuracy, achieving an average skip rate of 74.79%, yield rate of 99.95%, and recall of 80.20%. This indicates that when implemented in a real production line, the AI model can potentially reduce the need for up to two-thirds of aging facilities while maintaining a defect rate of less than 0.05%. Additionally, the resources saved can be focused on aging samples that require longer aging times, thereby reducing the rate of defective product release and enhancing product quality.

Author Keywords

LCM, Aging Line Defects, Imbalance Learning, Machine Learning, Cost-sensitive Learning, Smote ENN.

1. Introduction

In semiconductor display manufacturing, aging refers to the controlled testing process applied to devices, particularly LCDs (liquid crystal displays) and OLED (organic light-emitting diode) displays. This process simulates long-term usage to assess the devices' longevity and performance over time. The process of module (MOD) aging involves subjecting LCM products to a high-temperature environment and utilizing testing equipment to capture a range of visual images. Subsequently, visual image signals are utilized to evaluate the quality of LCD panels, ensuring that any defects are revealed. High-temperature aging typically consists of three stages: the early failure phase, the mid-stage random failure phase, and the late-stage wear-out failure phase. These stages collectively form the "failure curve principle." The primary objective of the aging process is to detect early-stage failures. Common aging defects can be categorized into four types: 1) line defects, 2) bubble defects, 3) mura defects, and 4) electrical defects. Each display panel undergoes an aging test lasting approximately 1-2 hours. For some high-end products, a full inspection of each panel is required to ensure quality control. However, with advancements in manufacturing technology, defect rates have significantly decreased, often as low as 0.1%. Conducting full aging inspections under such circumstances leads to substantial waste of time, production capacity, and energy resources.

To address this issue, this paper incorporates ML (machine learning) algorithms to predict the aging quality of each panel using real production data. This approach enables precise control by allowing defect-free panels to bypass additional testing while dynamically adjusting the aging duration for high-risk panels based on their risk levels. However, the defect rate in aging data is extremely low, resulting in an imbalanced data distribution. Consequently, the model's predictions tend to favor the majority class [1]. To mitigate this challenge, various approaches have been proposed, broadly categorized into two main types [1]. The first category involves data sampling techniques, such as SMOTE algorithms, which resample data from different classes to create a more balanced distribution. The second category focuses on algorithmic modifications, which adapt base learning methods to better handle minority class classification. For instance, Wei et al. [2] proposed a weighted sampling approach that balances datasets by filtering the majority class data based on their weighted complexity, retaining only the most representative samples to achieve a balanced distribution. Tanimoto et al. [3] utilized cost-sensitive learning combined with stratified sampling for imbalanced classification tasks. They enhanced baseline SVR and logistic regression models by incorporating near-miss positive instances, achieving a balanced accuracy exceeding 90%. Similarly, UmaRani et al. [4] introduced a hybrid method combining Random Forest classification with grey wolf meta-heuristic optimization for credit card fraud detection. Their results showed an improvement in accuracy from 0.87 to 0.946, outperforming conventional machine learning algorithms.

In this paper, a hybrid machine learning algorithm combining cost-sensitive learning and SMOTE-ENN techniques is proposed to classify imbalanced LCM aging data. The algorithm demonstrates strong performance in classifying the data with a limited number of defective samples. The AI model has been integrated into MOD production lines, where it automatically classifies defective products with high accuracy. The following chapters will provide a detailed explanation of the methods used for training and selecting the machine learning model.

2. LCM Data Collection

Data on LCM aging defects, collected over a six-month period, shows that 60% of defective products are due to particles originating in the array production process. These particles disperse across various module regions and are detectable by AOI (Automated Optical Inspection) and AOH (Automated Optical Height) systems. To train a machine learning model, we only collected features strongly correlated with MOD aging quality. Specifically, key AOI parameters are used as features, including total defect count, defect counts in high-risk regions, specific defect codes, and flags indicating if anomalies were detected by imaging or required manual repair.

To improve feature selection, the data is further segmented by particle size and grayscale levels detected by AOI. Additionally, certain regions, such as Levels 5 and 6 shown in Table 1, are designated as high-risk zones for aging line defects. Finally, the YMS (Yield Management System) is employed to extract over 100,000 historical records from aging line data.

ISSN 0097-996X/25/5602-1028-$1.00 © 2025

Table 1. The definitions of levels and zones of MOD aging line data.

Level	Zone / Type	Description
10	Fanout Region	Fanout Region
9	Golden Inspection	Golden Inspection
8	GOA Periodic Inspection	GOA Periodic Inspection
7	Special Flag	Special Flag
6	Cross	GD/CD Cross，High Risk Region
5	TFT	Big TFT + Small TFT + Share Bar/gate Cross，High Risk Region
4	GATE	Non-high-risk Region, Gate/COM
3	Data	Non-cross Data Line
2	Share bar	Non-cross Share Bar Line
1	Background	Open Region

3. Model Training and Validation Methodology

The flow chart for training and validating imbalance classification ML model is shown in Fig. 1. Feature engineering is firstly conducted for raw MOD aging dataset. The technique of label encoding is adopted for converting categorical variables into numerical format. Particularly, target variable of majority (non-defective aging) is encoded as 0, while that of minority (defective aging) is encoded as 1. Log transformation technique is used to stabilize variance, make the data more normal distribution-like, and improve the interpretability of the results. Subsequently, the mutual information algorithm is used to select the features that exhibit the highest correlation with the target variable.

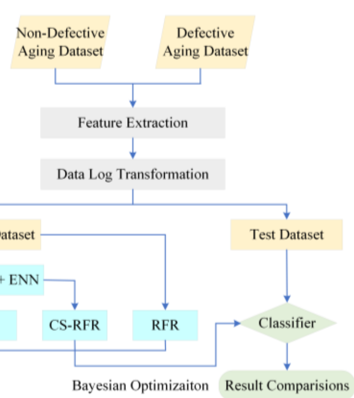


Figure 1. The flowchart for training and validating imbalance classification machine learning model.

Resampling techniques such as SMOTE, ADASYN, and ENN are utilized to tackle class imbalance by generating synthetic samples for the minority class. This approach helps the model learn more effectively from underrepresented data, ultimately enhancing overall classification performance. To enhance the generalization performance of the ML model, 8-fold cross-validation and stratified shuffle split techniques are employed during training. The algorithms XGBoost, Random Forest, LightGBM, and ExtraTrees are compared, and their hyperparameters are optimized using Bayesian optimization algorithm. Cost-sensitive learning method is employed to further improve the accuracy of minority class classification.

3.1 Data preprocessing

The labeling of MOD aging data presents a highly imbalanced binary classification challenge, with a ratio of non-defective to defective products of approximately 250:1. To better visualize the dataset distribution, t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to reduce the dimensionality of the data features. As shown in Fig. 2, the minority samples (defective aging products), depicted in blue, are widely scattered and almost completely overlap with the distributions of the majority class (non-defective aging products). More critically, there is no clear boundary separating the distributions of the majority and minority class samples.

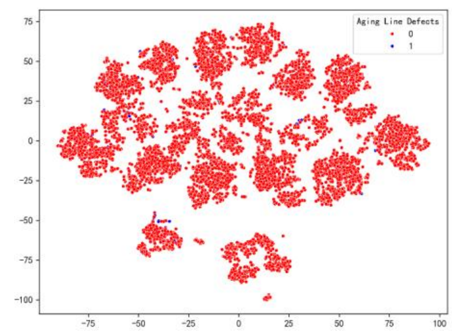


Figure 2. The distributions of no-defective and defective aging datasets.

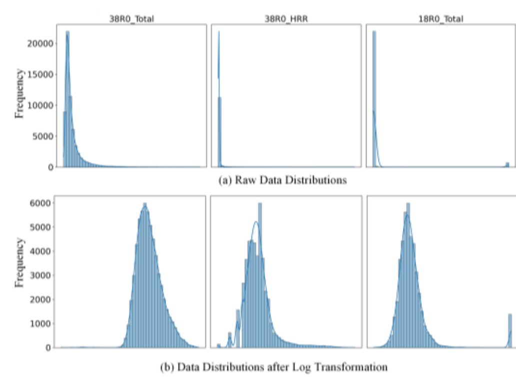


Figure 3. The features with long-tail distribution are converted to normal distribution by using log transformation.

To balance the dataset, hybrid approaches incorporating sampling algorithms such as SMOTE, ADASYN, Tomek Links, and ENN are utilized. Since many features exhibit a long-tail distribution, log transformation is applied to preprocess these features. As illustrated in Fig. 3, features initially following a long-tail distribution are effectively transformed into a normal distribution after log transformation. The equation for the log transformation is provided in Eq. (1):

$$Y' = \log(Y + c) \tag{1}$$

where Y represents the original feature value, c is a constant added to prevent taking the logarithm of zero (set to 1 in this study), and Y' is the transformed value.

3.2 Metrics for evaluating AI model

The choice of metrics in imbalance learning is crucial because traditional metrics like accuracy can be misleading when evaluating models on imbalanced datasets. In this paper, metrics such as F1-score, PR-AUC, recall, and G-mean are employed to evaluate the overall performance of machine learning models, with an emphasis on the accuracy of minority class classification. PR-AUC stands for Precision-Recall Area Under the Curve. It is a performance metric used to evaluate the effectiveness of a binary classification model, particularly in scenarios where the classes are imbalanced. The equations of F1-score, recall and G-mean are specified in Eq. (2), Eq. (3) and Eq. (4):

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$G - mean = [(TP/(TP + FN) \times (TN/TN + FP)]^2 \quad (4)$$

4. Results and Discussions

To identify the most suitable ML model for classifying aging products, various algorithms are compared. The hyperparameters of each model are optimized using Bayesian optimization and 8-fold stratified cross-validation. The mean and variance of each performance metric are calculated to evaluate the models. Additionally, ML algorithms are compared both with and without cost-sensitive learning methods.

As shown in Table 2, without applying cost-sensitive learning methods, XGBoost and LightGBM outperform Random Forest and ExtraTrees classifiers in overall performance. For standard ML algorithms, the mean Recall and G-Mean scores of the XGBoost model are 0.527 and 0.713, respectively, compared to 0.307 and 0.545 for the ExtraTrees classifier. Interestingly, cost-sensitive learning does not significantly benefit XGBoost and LightGBM models but leads to notable improvements in the evaluated metrics of Random Forest and ExtraTrees models. For instance, the mean G-Mean score of Random Forest increases from 0.616 to 0.729 after employing cost-sensitive learning.

To prioritize minority class accuracy, metrics such as Recall and G-Mean are given greater weight when selecting the optimal model. In this study, the Random Forest model with cost-sensitive learning is chosen, as it achieves relatively high Recall and G-Mean scores without excessively compromising the classification accuracy of the majority class.

Table 2. Comparisons of different machine learning algorithms on defective aging product classification.

No.	Model Name	Metric	Score (Normal)	Score (CS)
1	XGB	F1	0.605±0.089	0.605±0.089
		Recall	0.539±0.126	0.539±0.126
		PR_AUC	0.648±0.092	0.648±0.092
		G-Mean	0.716±0.081	0.716±0.081
2	LGBM	F1	0.62±0.093	0.62±0.093
		Recall	0.527±0.123	0.527±0.123
		PR_AUC	0.662±0.094	0.662±0.094
		G-Mean	0.713±0.082	0.713±0.082
3	RF	F1	0.517±0.11	0.543±0.122
		Recall	0.393±0.096	0.602±0.11
		PR_AUC	0.65±0.096	0.647±0.089
		G-Mean	0.616±0.081	0.729±0.061
4	ETC	F1	0.437±0.086	0.499±0.194
		Recall	0.307 ±0.069	0.733±0.069
		PR_AUC	0.599±0.134	0.586±0.16
		G-Mean	0.545±0.07	0.727±0.119

Experiments were also conducted using various resampling methods. Table 3 summarizes the performance of the Random Forest model with cost-sensitive learning under different sampling techniques. The results indicate that resampling methods significantly influence the classification accuracy of the model. Specifically, ENN with a ratio of 6.4:1 and the hybrid method combining ADASYN and ENN with a ratio of 4.3:1 achieve considerably higher Recall and G-Mean scores compared to other approaches. In this study, ENN is selected as the preferred method, as it delivers higher F1 and PR_AUC scores while maintaining Recall and G-Mean scores comparable to the hybrid method of ADASYN and ENN.

Table 3. Performance summary of Random Forest with various resampling methods.

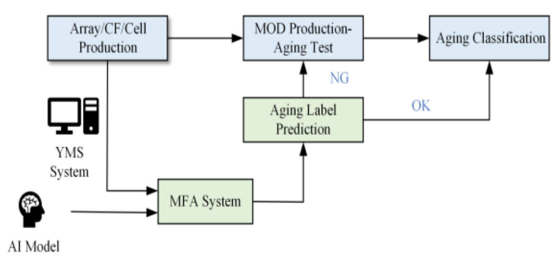
Resampling Method	Ratio	F1	PR_AUC	Recall	G-Mean
Blank	6.7:1	0.644	0.664	0.674	0.798
ENN	6.4:1	0.593	0.627	0.712	0.804
Smote	5:1	0.631	0.652	0.655	0.786
	2:1	0.617	0.641	0.599	0.756
	1:1	0.614	0.641	0.577	0.744
Smote +Tomek	5.4:1	0.62	0.642	0.657	0.785
	2.1:1	0.617	0.641	0.605	0.759
	1:1	0.618	0.645	0.585	0.749
Smote+ ENN	4.7:1	0.577	0.613	0.701	0.796
	1.7:1	0.575	0.607	0.676	0.784
	1.2:2	0.562	0.595	0.66	0.775
ADASYN	4.8:1	0.622	0.644	0.669	0.792
	2:1	0.612	0.634	0.629	0.771
	1:1	0.603	0.626	0.606	0.758
ADASYN +Tomek	5.2:1	0.619	0.643	0.679	0.795
	2:1	0.607	0.629	0.628	0.769
	1:1	0.607	0.63	0.608	0.759
ADASYN + ENN	4.3:1	0.564	0.61	0.728	0.804
	1.6:1	0.552	0.599	0.714	0.795
	1.3:1	0.541	0.587	0.696	0.784

Table 4 presents the prediction performance of the Random Forest model, incorporating cost-sensitive learning and ENN resampling techniques. The model was tested on 210768 pieces of panel collected from a real MOD production line. Three evaluation metrics, skip ratio, yield rate, and recall, were used to assess model performance. The skip ratio is defined as the proportion of predicted non-defective data relative to the total data size. For the machine learning model to be implemented in the real MOD aging product line, the following accuracy thresholds shoud be met: skip ratio ≥ 66.7%, yield rate ≥99.95%, and recall ≥60%. From the table, it is evident that the model's prediction accuracy is exceptionally high, with average scores of 74.79% for skip ratio, 99.95% for yield rate, and 80.20% for recall. This suggests that, if implemented in the real MOD production line, the ML model could reduce the need for approximately two-thirds of the aging inspection facilities, while maintaining a defective product ratio of less than 0.05%.

Table 4. The prediction performance of the Random Forest model incorporating cost-sensitive learning and ENN resampling techniques.

Confusion Matrix		**Prediction**		
		0-OK	1-NG	Skip Ratio/ Recall
Actual measurement	0-OK	157552	52817	**74.9%**
	1-NG	79	320	**80.20%**
	Yield rate	**99.95%**		

The ML model has been successfully deployed in real LCM production lines. Figure 4 depicts the flowchart of the ML model's automated classification of LCM aging data. A YMS system collects raw data from the array, CF, and cell production lines. This raw data is then fed into the MFA system, where the trained ML model is integrated. The data is subsequently processed, and the ML model predicts the aging quality, assigning a label to each dataset. Products labeled as "OK=AGSK" are sent directly to the next process, while those labeled as "NG=null" are routed to facilities for aging tests. This approach significantly enhances the efficiency and quality of aging classification, reducing the reliance on aging test facilities by approximately two-thirds.



(a) The flowchart of ML model automatically classifying LCM aging data.

(b) The window displaying the predicted results in MOD production lines.

Figure 4. (a) The flowchart of ML model automatically classifying LCM aging data; (b) The window displaying the predicted results in MOD production lines.

5. Conclusions

This paper presents a machine learning-based solution to tackle the imbalanced class problem in semiconductor display manufacturing. First, a log transformation is applied to normalize the raw data, converting its long-tailed distribution into a relatively normal distribution. Next, several machine learning algorithms are evaluated based on their prediction accuracy, with hyperparameters optimized using Bayesian optimization. The best-performing algorithm is then selected as the baseline model and further enhanced through the integration of cost-sensitive learning and SMOTE resampling techniques.

In this study, the Random Forest algorithm, combined with cost-sensitive learning and ENN sampling techniques, proves to be the optimal choice for semiconductor display aging classification. The model is tested on 210768 pieces of panel real-world production datasets, achieving an average skip ratio of 74.79%, a yield rate of 99.95%, and a recall of 80.20%. These results demonstrate that the ML model can potentially reduce the need for up to two-thirds of aging facilities while maintaining a defect rate as low as 0.05% when implemented in real production lines. Additionally, the resources saved can be focused on aging samples that require longer aging times, thereby reducing the rate of defective product release and enhancing product quality.

6. Acknowledgements

The authors would like to thank the team at Getech Co., Ltd. for their support in implementing the ML model into our LCD production lines.

7. References

1. Sun Y. M., Wong A. K. C., & Kamel M. S., Classification of imbalanced data: a review, International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4), 687-719.
2. Wei W., Jiang F., Yu X. Du J., An under-sampling algorithm based on weighted complexity and its application in software defect prediction. 5th International Conference on Software Engineering and Information Management (ICSIM), 2022, 38-44.
3. Tanimoto A., Yamada S., Takenouchi T., Sugiyama M., Kashima H., Improving imbalanced classification using near-miss instances. Expert Systems with Applications, 2022, 201, 117130.
4. UmaRani, V., Saravanan, V., & Tamilselvi, J. J., A hybrid grey wolf-meta heuristic optimization and random forest classifier for handling imbalanced credit card fraud data. Int J Intell Syst Appl Eng, 2023, 11(9), 718-734.