New statistical analysis methodology to forecast the memory cell behavior before reliability test

S. Perrin, V. Della Marca, T. Kempf, Marc Bocquet, L. Welter, J.M. Moragues, A. Regnier, J.M. Portal

New statistical analysis methodology to forecast the memory cell behavior before reliability test

S. Perrin[a,b*], V. Della Marca[b], T. Kempf[a], M. Bocquet[b], L. Welter[a], J. M. Moragues[a], A. Regnier[a], J. M. Portal[b]

[a] *STMicroelectronics, 190 avenue Celestin Coq, 13106 Rousset, France*
[b] *Aix-Marseille University, IM2NP, CNRS, UMR 7334, 5 rue Enrico Fermi, 13453 Marseille, France*
*sebastien.perrin01@st.com

Abstract

In this paper, a machine learning method is proposed implementing the Principal Component Analysis to study the statistical EEPROM endurance degradation. This technique is firstly applied to an UV irradiated memory array. Then, the Density Based Spatial Clustering of Applications with Noise and the Gaussian Mixture Model are presented to extract the minority population of cells. The reliability test study demonstrated the ability of the proposed technique to correlate electrical parameters to forecast the quality and performance of a memory array. Compared to the classical threshold voltage (V_{th}) analysis, this method is more effective for predicting which population will experience greater degradation.

1. Introduction

The maturity ramp-up is a crucial step in the industrialization and commercialization of eNVM technology. To achieve this, several milestones are validated through high statistical electrical tests on various device structures to assess their performance.

Firstly, large amounts of single cells are subjected to Wafer Level Reliability (WLR) to investigate intrinsic cell characteristics. However, WLR requires many wafer resources. Parametrical Tests (PT) are also performed on single-cell structures to measure mainly electrical parameters such as threshold voltage (V_{th}). Secondly, Cell Array Stress Test [1] (CAST) allows for bit failure detection on the same memory array, but it cannot provide the position and number of bits that fail. Finally, the test chip, which is a product-like structure, provides high statistical data and detects the weak bit cells (extrinsic population) position and number [2,3]. However, this structure, remaining complex to test and expensive to design, is normally used for mature technologies.

A fully addressable EEPROM memory array was developed in [4], called SuperCAST. This structure supports endurance and retention test, as well as cell selection flexibility. In other words, it is possible to address the memory array as a CAST or measure the whole current-voltage characteristics of each single cell, along with their position in the memory array. As a result, various electrical parameters can be extracted. Therefore, it is important to develop tools to analyze the measured characteristics to extract meaningful insights from these statistical data.

In this paper, we propose the Principal Component Analysis (PCA) method to analyse our statistical data obtained with the SuperCAST. This tool can reduce the dimensionality of the dataset by calculating principal components from all features [5], making it suitable for the machine learning domain. As a result, better data visualisation is available, and we can identify how features describe each observation in the dataset. The aim of this paper is to determine relationships between EEPROM cells at fresh state and after endurance up to 10^6 cycles. It is important to notice that this kind of reliability analysis is independent from the studied device (Flash, ReRAM, or transistors). The main advantage is that it can be employed at the early development stage of new technology, to save time and costs.

2. Principal component analysis methodology

In this second section, the PCA flow is described with a dataset obtained by 761 random selected cells in the UV irradiated SuperCAST memory array. The PCA method involves four preliminary steps: i) A dataset is created by measuring the drain current (I_d) versus the control-gate voltage (V_{cg}) characteristic of each single cell in the memory array (read operation). The extracted features are the threshold voltage at I_d=100nA (V_{th}), the peak drain current (I_d), the peak of transconductance (gm_{max}), and the subthreshold slope (Ss). ii) The electrical parameters are normalized using the preprocessing phase [5]. Moreover, iii) the PCA process calculates the covariance matrix, iv) which is used to obtain the eigenvalues and eigenvectors. The latter are defined as principal components [5,6]. In addition, the variance explained ratio (V_r) of a principal component

is calculated as the ratio of its corresponding eigenvalue to the sum of all eigenvalues. Hence, V_r indicates the proportion of the total variance in the dataset, explained by that principal component (Fig. 1.a). As a result, principal components are chosen so that the cumulated variance explained ratio is bigger than 80% (Fig. 1.b). Concerning this criterion, in Fig. 1.c each point corresponds to one bit cell in the population, and the data are projected onto the principal component. We observe that the data projection is homogeneous and represents the intrinsic behavior of UV irradiated EEPROM memory array. In Fig. 1.d, the circle of correlation is plotted to visualize the relationships between each feature. The electrical parameters are represented by oriented vectors. Colinear vectors indicate strong correlation between the parameters, while orthogonal vectors suggest their independence. It is important to notice the red dots give the direction of the higher value of the feature. In the next section, we apply PCA to an existing dataset of extrinsic population that were artificially generated using the SuperCAST vehicle. Our goal is to assess its capability to automatically detect different groups of populations and to introduce machine learning for reliability studies.

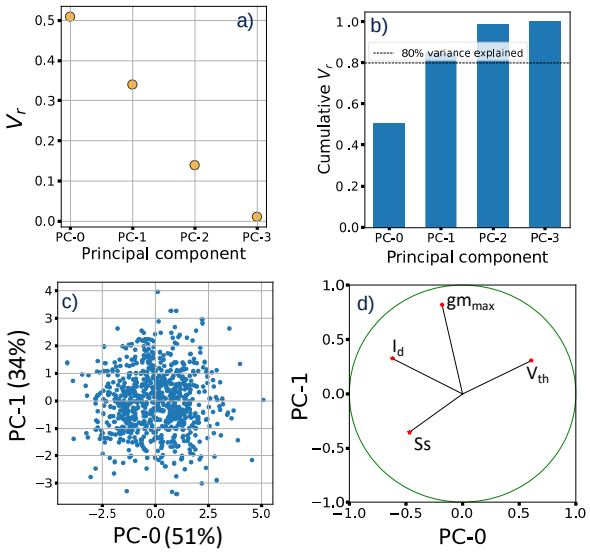

Fig. 1. a) The variance explained ratio for each principal component. b) The cumulative explained variance, as a function of principal components needed to reduce dimensionality while minimizing information loss. c) PCA: two-dimensional data projection of the chosen principal components (PC-0 and PC-1) that contains the most important information. d) The circle of correlation: features relationship.

3. Principal component analysis assessment

Thanks to the SuperCAST, we can program or erase any group of cells in the memory array. To assess the PCA methodology, all 761 selected cells are erased in the memory array (high V_{th}), considered then as the intrinsic population while only one wordline (128 cells) is programmed (low V_{th}), considered as the extrinsic population.

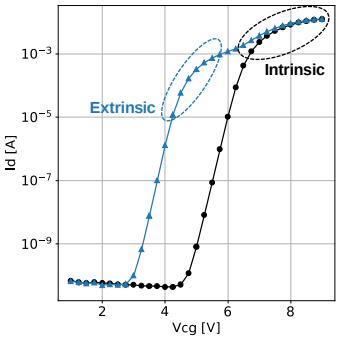

Fig. 2. Id-Vcg of SuperCAST, the black line represents the 761 erased memory cells, and the blue line is the characteristic where a word line has been programmed.

In Fig. 2, the typical characteristics of a simple CAST structure under the previously described conditions are shown. Therefore, knowing the position and the number of extrinsic cells, it will be easy to validate the PCA functioning. To accomplish this, the PCA has been applied to dataset containing only one extrinsic wordline data and the corresponding results are depicted in Fig. 3. It is important to note that the cumulative variance explained ratio reached 80% for two principal components, PC-0 and PC-1. As a result, the two main groups are highlighted on the principal component space (Fig. 3.a). The cell indexing is used to verify that the minority population corresponds to the 128 cells of the programmed wordline in the SuperCAST. The opposite orientation of V_{th} and I_d on the circle of correlation (see Fig. 3.b) indicates that cells with low V_{th} exhibit high I_d and vice versa. Moreover, their length suggests that they contribute significantly to the PC-0 value. Inversely, gm_{max} mainly impacts the PC-1 value. The Ss parameter presents a weak projection on both principal components meaning that its correlation with the other parameters is low. Finally, according with the circle of correlation, the minority population represents lower V_{th} cells characteristics, that is consistent with our experiment. Thus, the PCA can be used to automatically detect two clusters.

In our example the two groups of cells are well distinguished and visible. So, we implemented two clustering methods for comparison: Density-Based

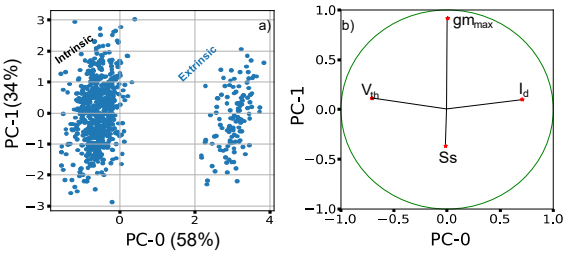

Fig. 3. PCA application on dataset with artificially generated extrinsic population. a) Data projection on principal component space. b) Features relationship in the circle of correlation plot.

Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Model (GMM).

4. Clustering methods

Firstly, the DBSCAN, that is an unsupervised method, makes clusters by grouping at least n points at a distance lower than ε from a defined core point, when a core point is within the epsilon-neighborhood of another core point, then they belong to the same cluster [7,12]. In addition, the minimum sample parameter must be specified to define the minimum number of points that is used to calculate ε. The minimum sample depends on the dataset dimension [8]. When a core point is within the epsilon-neighborhood of another core point, then they belong to the same cluster. We plotted in Fig. 4.a the calculated ε as a function of the data points sorted by distance. This plot is obtained by calculating, for each point in the principal component space, the average distance between its n nearest neighbor. The curve obtained in Fig 4.a is subtracted from the maximum value of the average distance (ε_{max}). The value of the result is taken and multiplied by the values of the x-axis (Data points). The result of this transformation is plotted against the "Data points" to obtain a bell curve in Fig 4.b, whose maximum value gives us the threshold limit (ε_{th}). Thus, the algorithm can identify the clustered population, while the remaining data is defined as noise (red region) [7].

Secondly, the GMM is a probabilistic model that assumes the dataset is generated by linear combination of Gaussian populations [9]. Moreover, each Gaussian function is a component with its own mean and covariance. To identify clusters, the user specifies the number of components, and an Expectation-Maximization algorithm is executed to calculate the probability of each point in the principal component space belonging to a component. This illustrates that the GMM is a supervised method.

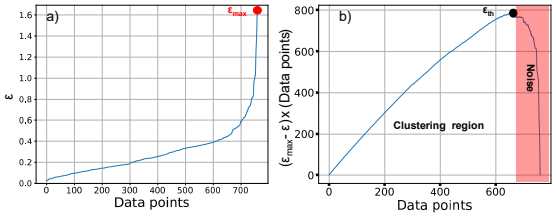

Fig. 4. a) The epsilon (ε) graph as a function of sorted data points. b) The latter is transformed so that (ε_{max}-ε) * (Data points) to obtain the epsilon threshold for clustering (ε_{th}).

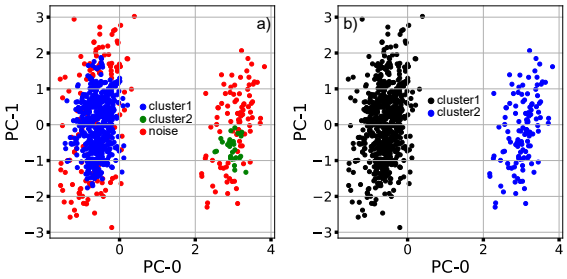

Fig. 5. Clustering method comparison: a) DBSCAN and b) GMM clustering.

Fig. 5 illustrates the comparison between the DBSCAN and GMM clustering methods. The DBSCAN clustering algorithm found three groups in respect to the calculated epsilon value. In Fig. 5.a, the blue and the green populations are considered as clusters because the mean distance between core point in each group is below the epsilon threshold. Inversely, the red group is considered as noise. Moreover, the DBSCAN can detect the majority (cluster1) and minority (cluster2) populations.

In the case of GMM, the number of components was manually set to 2. As shown in Fig. 5.b, the algorithm clearly identified the majority and the minority populations (cluster1 and cluster2 respectively). In this case, the technique requires human intervention to detect obvious groups in principal component space. However, the GMM method can become an unsupervised tool by implementing Bayesian Information Criterion (BIC) into the algorithm to automatically define the number of components [10]. Both clustering methods must be adapted to the provided data and can infer the number of clusters. In section 5, we applied PCA comparing DBSCAN and GMM clustering method on a reliability dataset obtained with the EEPROM SuperCAST.

5. Reliability test and PCA

We defined a classical experimental endurance protocol to show a possible application of the PCA using both previous clustering methods. The aim of this study is to find a relationship between the

electrical parameter variation during reliability test and their initial electrical properties. The cells have been programmed and erased by Fowler-Nordheim tunnelling and read at the beginning and after 10^6 programming/erasing cycles. The endurance test has been carried out connecting all the memory devices in parallel, while the reading has been performed cell by cell. From the measurements, a dataset is created (as depicted in section 2) for both erased states before and

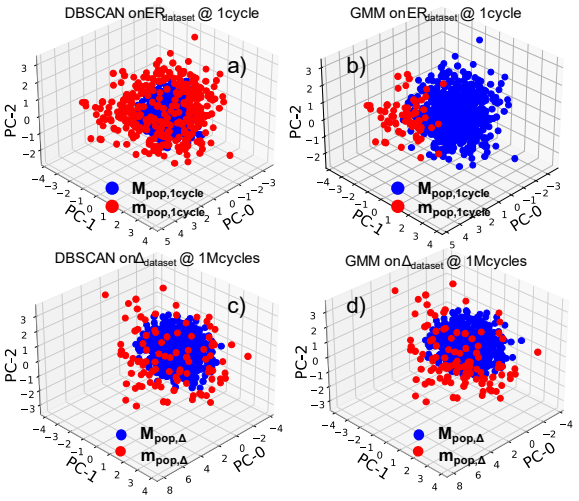

Fig. 6.a) The PCA with DBSCAN and b) PCA with GMM clustering calculated for 1cycle erased state dataset, and the same representation using the delta dataset c) with DBSCAN and d) with the GMM clustering methods.

after stress. Moreover, we calculated the delta dataset ($\Delta_{dataset}$) which is the difference between each electrical parameter after and before stress. In general, the extrinsic population is visible at low V_{th}, but the presence of minority (m_{pop}) and majority (M_{pop}) population is explored to prevent reliability issues in the memory array. Applying the PCA to the $ER_{dataset}$ at 1 cycle, the cumulative variance explained ratio reached 80% with three principal components. This implies that the PCA clustering results should be plotted in a three-dimensional principal component space as reported in Fig. 6.a and 6.b for DBSCAN and GMM clustering, respectively. One can notice that in this case, the cell population is homogeneous, and the DBSCAN method automatically detects majority and minority population ($M_{pop,1cycle}$=463, $m_{pop,1cycle}$=298 respectively), while for GMM, specifying the number of clusters is required ($M_{pop,1cycle}$=688, $m_{pop,1cycle}$=73). In Fig. 6.c and 6.d the PCA is performed on the $\Delta_{dataset}$ and the two clustering methods are again compared. It is worth to notice the displacement of $m_{pop,\Delta}$. In both cases, the minority population is located around the majority population ($M_{pop,\Delta}$) core, and almost the same number of cells is detected. The results are reported in Table 1. To compare the DBSCAN and

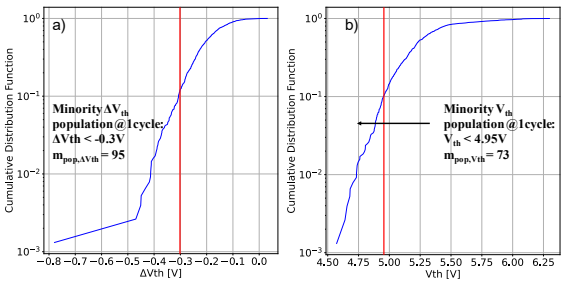

Fig. 7.a) Cumulative threshold voltage variation (ΔV_{th}). The population which has variation lower than -300mV are defined as most degraded memory and represents 10% of the most degraded population. b) Cumulative threshold voltage distribution before stress. The minority population is defined as Vth smaller than 4.95V, representing 10% of the smaller threshold voltage memory bit cells at t_0.

GMM clustering methods, the selectivity (true negative rate) and the sensitivity (true positive rate) criteria are assessed. On the one hand, the selectivity defined as the probability of $m_{pop,1cycle}$ and $m_{pop,\Delta}$ intersect. This probability is the ratio between the population intersection and the population of the most degraded memory bit cells. This ratio provides insights into the proportion of most degraded device population that can be in $m_{pop,1cycle}$. On the other hand, the sensitivity defined as the probability of $M_{pop,\Delta}$ intersects with $M_{pop,1cycle}$ is calculated by the ratio between the majority population intersection relative to $M_{pop,\Delta}$. This second ratio corresponds to the proportion of the non-degraded devices observed after stress ($M_{pop,\Delta}$) in the initial majority population ($M_{pop,1cycle}$). For example, concerning the DBSCAN clustering method, Sensitivity=60% means that 60% of the $m_{pop,\Delta}$ (population of 95 memory bit cells that will be more impacted by the degradation) belongs to the initial minority population $m_{pop,1cycle}$ (population of 298 memory bit cell detected before stress). In addition, Selectivity= 64% means that the non-degraded population observed after stress ($M_{pop,\Delta}$) corresponds to 64% of the initial majority population ($M_{pop,1cycle}$). This result means that most of the weaker memory bit cells can be detected in a minority population before reliability test, but that minority population is mainly composed by non-degraded memory bit cells.

Additionally, the classical V_{th} analysis is assessed considering the criteria defined in Fig. 7.a. Indeed, the classical method consists of assessing the fraction of the most degraded memory bit cells ($\Delta V_{th} \leq$ -300mV) detected in the minority population where V_{th} is smaller than 4.95V at 1cycle, as depicted in Fig. 7.b. The population sizes are taken such that they represent 10% of their respective population and the selectivity

and sensitivity are revealed in Table I. Thus, a population of the 76 most degraded V_{th} cells ($m_{pop,\Delta Vth}$) is selected and only 6% (sensitivity) of them are included in the initial $m_{pop,Vth}$ of 76 lowest V_{th} values. Moreover, 89% (selectivity) of the non-degraded memories are correctly identified in the majority population at t_0. Thus, the classical method detects few degraded V_{th} bit cells at t_0 but detects correctly the non-degraded population at t_0. The significant selectivity observed can be attributed to the larger population size of the high V_{th} group, which is substantially greater than that of the low V_{th} group. So, it is easier to predict the non-degraded population at t_0 with the classical Vth analysis. The sensitivity and selectivity are also computed for DBSCAN and GMM methods.

Table 1
Results of data correlation between most degraded population and their electrical properties at t_0 for two clustering methods and a classical method

PCA (761 samples)	**DBSCAN**	**GMM**	**ΔV_{th}**
$m_{pop,1cycle}$ ($ER_{dataset}$)	298	73	76
$M_{pop,1cycle}$ ($ER_{dataset}$)	463	688	685
$m_{pop,\Delta}$ ($\Delta_{dataset}$)	95	110	76
$M_{pop,\Delta}$ ($\Delta_{dataset}$)	666	651	685
Sensitivity	**60%**	**38%**	**6%**
Selectivity	**64%**	**95%**	**89%**

In the case of GMM method, 38% of the weaker bit cells (110) are initially detected in the minority population (73). It means that for this dataset, GMM can detect only few weak bit cells and in spite of a bigger $m_{pop,\Delta}$ population, the sensitivity is smaller than DBSCAN. In addition, the DBSCAN method shows that 60% of the weaker bit cells (95) are initially detected in the minority population (298), meaning that most of the weaker bit cells are detected at t_0, but they correspond to a small part of the $m_{pop,1cycle}$.
Therefore, according to Table 1, the weak memory bit cell detection at t_0 is better by considering several electrical parameters (DBSCAN and GMM) rather than only one electrical parameter (V_{th}). In the case of the selectivity criteria, the GMM method shows that 95% of the non-degraded population observed after stress ($M_{pop,\Delta}$) is also in the initial majority population while the DBSCAN method provides 64%. Therefore, according to both criteria, it is evident that the classical method is the worst example to detect the weaker bit cells at t_0. In the case of DBSCAN and GMM, it is difficult to determine which one forecasts better the weak bit cells population at t_0. So, it is important to find a trade-off between these two criteria and select the most appropriated method in respect to the issue we face. For example, if the objective is to assess a process flow occurred as expected, it could be interesting to focus on the potential memory bit cells failure at t_0 (sensitivity) capability aiming to save reliability testing time. In this way, the DBSCAN method appears to be more suitable to predict the non-degraded population in respect to GMM.
For this potential application, the sensitivity stability of these methodologies is assessed for various dataset size and various population dispersion randomly selected from the t_0 main 4000-sample dataset and delta dataset. The procedure is depicted in Fig. 8.
Firstly, the stability is studied for five dataset sizes: 761, 1000, 1500, 2000, 3500 (step 1). To create those datasets, a random draw is performed from both main datasets (step 2). For example, 761 samples are randomly selected from the main dataset (4000-sample) and the same population is selected from the delta dataset to create two smaller sub-datasets. So, a new 761-size initial dataset ($ER_{dataset,1cycle}$) and delta dataset ($\Delta_{dataset}$) are created. This procedure is repeated fifty times to generate fifty pairs of sub-datasets with the same population size. On the one hand, this step permits to save time measurement, and on the other hand, it permits to consider various random population dispersion.
Thus, for each sub-dataset pairs, both DBSCAN and GMM clustering methods and PCA are performed to measure the relationship between $m_{pop,\Delta}$ and $m_{pop,1cycle}$ (step 3). Consequently, the sensitivity is calculated considering both clustering for all sub-dataset pairs and for each size (step 4). Moreover, both methods are still compared with the classical one as described in Fig. 7. Finally, fifty sensitivity criterion is calculated for both clustering techniques in respect to the sampled population size (step 5).
The results are illustrated in Fig. 9 and Fig. 10 for GMM and DBSCAN respectively.

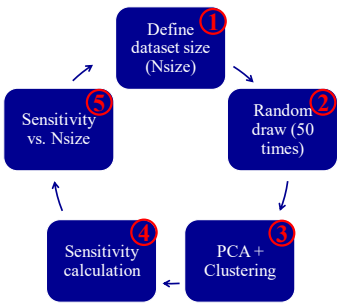

Fig. 8. Stability evaluation procedure of DBSCAN and GMM clustering methods.

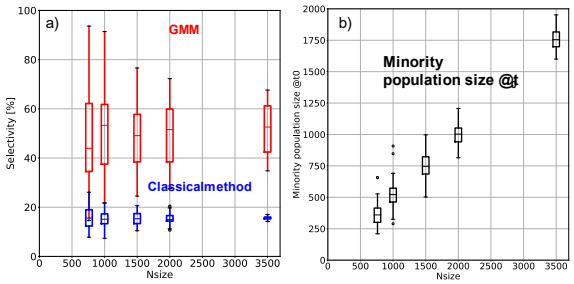

Fig. 9.a) Sensitivity boxplot from GMM clustering (red) vs. dataset size (Nsize) compared with classical method (blue) b) The minority population size detected with GMM from various $ER_{dataset}$ size.

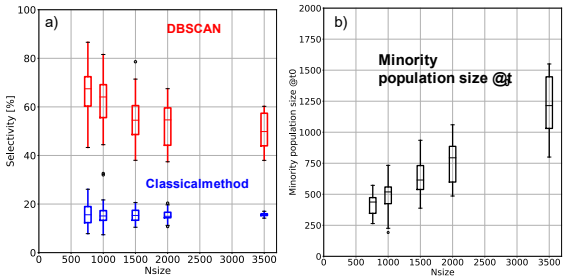

Fig. 10. a) Sensitivity boxplot from DBSCAN clustering (red) vs. dataset size (Nsize) compared with classical method (blue) b) The minority population size detected with DBSCAN from various $ER_{dataset}$ size.

The Fig. 9.a (resp Fig. 10.a), depicts the sensitivity boxplot (representing de fifty random draws) calculated from GMM (resp DBSCAN) method (in red) for various randomly sampled sub-dataset size. The sensitivity is also calculated with the classical method (in blue). The Fig. 9.b (resp Fig. 10.b) represents the corresponding minority population size ($m_{pop,1cycle}$) detected at t_0 thanks to the GMM (resp DBSCAN) technique.

In Fig. 9.a, the sensitivity median calculated with GMM is around 50% by increasing Nsize, while the sensitivity median from the classical method decreases from 30% to 20% by increasing Nsize. We can also notice the strong dispersion obtained with GMM, while the classical method results show small dispersion. For Nsize ≤ 2000, the GMM sensitivity boxplots encompass those obtained with the classical method. It means that for this dataset size, GMM can give the same prediction as that obtained by considering only the V_{th}. Moreover, Fig. 9.b illustrates that the median of the minority population at t_0 is the half of the dataset size whatever the Nsize. It means that one random draw over two conducts to a bigger $m_{pop,1cycle}$ size than the half of the sub-dataset. So, it cannot be possible to consider them as "minority population". Therefore, according to this results, GMM clustering seems to be statistically irrelevant to predict the weaker memory bit cells at t_0.

In Fig. 10.a, the sensitivity median calculated with DBSCAN decreases from 67% to 50% while the sensitivity median from the classical method decreases from 15% to 10% by increasing Nsize. Similarly, we can also notice that the sensitivity results obtained with DBSCAN are less dispersed (range of 25% to 40% standard deviation) than that obtained with GMM. In addition, the sensitivity boxplots do not encompass those obtained with the classical method, illustrating that DBSCAN clustering gives better sensitivity whatever the dataset size. Finally, the minority population size at t_0 represented in Fig. 10.b shows that for Nsize < 1500, the DBSCAN detects, for more than half of random draws, a bigger $m_{pop,1cycle}$ than the half of Nsize. So, in these cases, the sensitivity is not relevant for most of the random draw and the corresponding results cannot be considered for these Nsize. Similar observation is made for Nsize = 2000, except that less random draws have conducted into bigger $m_{pop,1cycle}$ than Nsize/2. Finally, for Nsize = 3500, the $m_{pop,1cycle}$ is smaller than Nsize/2 for each random sub-dataset. In this case, the sensitivity results are relevant to statistically predict weak memory bit cell population at t_0. Based on these findings, it can be concluded that, on average, 50% of the memory cells that are most prone to degradation can be identified within a minority population at t_0.

6. Conclusion

In this study we implemented a statistical analysis methodology to forecast statistically the memory bit cells degradation. The SuperCAST test vehicle combined with PCA are useful for large data collection and analysis. We presented a new way to use machine learning PCA methodology to assess reliability issues of a memory array. Thus, two clustering techniques, DBSCAN and GMM, were compared to find the best trade-off for the data analysis. In this study, GMM clustering is unsuitable to correlate degraded population in respect to initial electrical properties because groups are undiscernible. In addition, we demonstrated that is possible to achieve the correlation of many electrical parameters of memory devices extracted before and after an endurance test thanks to DBSCAN clustering. The stability of this method has been assessed by studying the variability of the results. Finally, the degradation prediction can be conducted by considering several electrical parameters rather than only one.

As PCA and unsupervised clustering technique can forecast weak memory bit cell in an initial minority population, they could be used for process manufacturing quality and qualification with reduced testing population to save time and costs.

References

[1] Capeletti P et al. Cast: An electrical stress test to monitor single bit failure. Microelectronics Reliability. vol 37, No 3, 1997, pp. 473-481, doi: 10.1016/0026-2714(95)00214-6.

[2] Kempf T et al. Threshold voltage bitmap analysis methodology: Application to a 512kB 40nm Flash memory test chip. International Reliability Physics Symposium, Burlingame, CA, USA, 2018, doi: 10.1109/IRPS.2018.8353642.

[3] Plantier J et al. Retention test and electrical stress correlation to anticipate EEPROM tunnel reliability issues. International Semiconductor Devices Research Symposium, College Park, MD, USA, 2009, doi: 10.1109/ISDRS.2009.5378306.

[4] Della Marca V et al. SuperCAST: a full free adressable memory array. International Conference on Microelectronic Test Structures, Cleveland, OH, USA, 2022, doi:10.1109/ICMTS50340.2022.9898189.

[5] Bro R and Smilde AK. Principal component analysis. Analytical Methods, vol 6, 2014, pp. 2812-2831.

[6] Bishop CM. "Continuous Latent Variables" in *Patter Recognition and Machine Learning*, Berlin, Germany, 2006, Springer, ch.12, sec.1, pp. 859-862.

[7] Ester M et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. International Conference on Knowledge Discovery and Data Mining, Portland OR, USA, 1996, pp. 226-231, doi: 10.5555/3001460.

[8] Sander J et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. Data Mining and Knowledge Discovery. vol 2, 1998, pp. 169-194.

[9] Sun W et al. Gaussian mixture model-based random search for continuous optimization via simulation. Winter Simulation Conference, Gothenburg, Sweden, 2018, doi: 10.1109/WSC.2018.8632380.

[10] Fraley C and Raftery A. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the Statistical Association. vol 97, 2002, pp. 611-631.

[11] Sumikawa N et al. Forward Prediction Based on Wafer Sort Data - A Case Study. International Test Conference, Anaheim, CA, USA 2011, doi: 10.1109/TEST.2011.6139174.

[12] Sumikawa N et al. Kernel based clustering for quality improvement and excursion detection. International Test Conference, Fort Worth, TX, USA, 2017, doi: 10.1109/TEST.2017.8242071