INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

Optimized CatBoost-Based Modeling for Accurate Semiconductor Yield Diagnosis and Tuning

KOPPISETTI VENKATA SURYA
Master Of Computer Applications
Ideal College Of Arts & Sciences,
Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

Mr. K. PRAVARDHAN
Assistant.Prof, Master Of Computer Applications
Ideal College Of Arts & Sciences,
Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

Dr. V.S.V. DEEPAK
HOD, department Of Computer science
Ideal College Of Arts & Sciences,
Autonomous, Affiliated To Adikavi Nannaya University - Rajamahendravaram
Kakinada

***Abstract*— Yield diagnosis and tuning in emerging semiconductor devices remains challenging due to limited experimental data and high fabrication costs during the research stage. Traditional machine learning models such as Support Vector Machine, Random Forest, and XGBoost provide reasonable predictions but often suffer from overfitting and poor handling of complex feature interactions. This work introduces an extension by integrating the CatBoost algorithm into the existing yield prediction framework. The proposed approach utilizes process parameters and electrical characteristics to model yield behavior and identify key influencing factors. Bayesian optimization is employed to iteratively refine model parameters and improve prediction performance. The extended model demonstrates improved accuracy with lower mean absolute error and higher explained variance compared to conventional methods. CatBoost effectively handles encoded features and data imbalance, leading to more stable predictions. This approach reduces the need for extensive experimentation and enables efficient yield optimization in early-stage semiconductor research environments.**

Keywords*— *CatBoost, Machine Learning, Yield Prediction, Semiconductor

I. INTRODUCTION

Semiconductor manufacturing is one of the most complex and precision-driven processes in modern engineering. As device dimensions continue to shrink and new materials such as emerging two-dimensional structures are introduced, maintaining high yield becomes increasingly difficult. Yield refers to the proportion of functional devices produced during fabrication, and even small variations in process parameters can lead to significant performance degradation. During the research stage, this challenge becomes more critical because experiments are limited, costly, and time-consuming.

Traditionally, yield improvement relies on extensive trial-and-error experimentation, where multiple process parameters are adjusted and evaluated through repeated fabrication cycles. This approach demands substantial resources, including materials, equipment usage, and skilled manpower. Moreover, early-stage semiconductor research often suffers from limited data availability, making it difficult to apply conventional statistical methods effectively. The lack of sufficient data also restricts the ability to identify root causes of yield loss with high confidence.

In recent years, data-driven techniques have gained attention as a means to improve decision-making in semiconductor manufacturing. Machine learning methods, in particular, offer the ability to analyze complex relationships between process parameters and device performance. These methods can extract patterns from available data and provide insights that are difficult to obtain through manual analysis. However, their effectiveness is often constrained by data scarcity, noise, and imbalance in experimental datasets.

Another key challenge lies in optimizing process parameters efficiently without performing a large number of physical experiments. Identifying the optimal combination of parameters requires intelligent search strategies that can balance exploration and exploitation. This is essential to reduce development time and cost while ensuring reliable device performance. Therefore, there is a strong need for advanced techniques that can handle limited data, improve prediction accuracy, and support efficient yield enhancement in semiconductor research environments.

II. RELATED WORK

Semiconductor yield improvement has been widely studied with a focus on understanding process behavior and minimizing defects during manufacturing. Early work by Tirkel (2013) introduced yield learning curve models, showing that yield improves gradually as process knowledge increases over time. This concept provided a structured way to estimate production performance and highlighted the importance of continuous learning in fabrication environments. In the same period, Chien et al. (2013) proposed data-driven fault detection and classification techniques to identify abnormal process conditions. Their work demonstrated that early detection of faults plays a key role in preventing defect propagation and improving overall yield.

Lenz and Barak (2013) extended the use of machine learning by applying support vector regression for virtual metrology.

Their approach reduced reliance on expensive physical measurements by predicting process outputs from existing data, enabling faster and more cost-effective monitoring. These early studies established the importance of data-driven approaches in semiconductor manufacturing. Moyne et al. (2016) further emphasized the role of big data and advanced process control systems, showing how large-scale data analytics can improve decision-making and yield optimization in complex manufacturing environments.

As semiconductor technologies advanced, new challenges emerged due to device scaling and material innovations. Cao et al. (2015) highlighted the increased variability and sensitivity in sub-10 nm and 2D semiconductor devices, which significantly complicate yield prediction. Similarly, Feng et al. (2017) analyzed advanced semiconductor architectures and showed that modern device designs introduce additional process variations, making traditional modeling techniques less effective. To address these challenges, Chen (2017) proposed artificial neural network models that can capture nonlinear relationships between process parameters and yield. This work demonstrated improved prediction accuracy compared to conventional statistical methods.

More recent studies have focused on integrating advanced artificial intelligence techniques for yield prediction. Stich et al. (2020) introduced an AI-based cascading classification system that combines multiple models to improve prediction performance. Their results showed that hybrid approaches are more effective in handling complex and high-dimensional manufacturing data. He (2022) discussed the rapid evolution of semiconductor manufacturing processes and emphasized the need for intelligent and adaptive systems to manage increasing complexity and variability.

Xu et al. (2022) proposed adaptive virtual metrology techniques for yield prediction in multi-batch wafer production. Their approach uses data-driven models that can adjust to process variations in real time, improving prediction reliability across different production conditions. This represents a shift toward dynamic and flexible modeling strategies that can handle changing manufacturing environments.

Overall, the literature shows a clear progression from statistical and rule-based methods to advanced machine learning and AI-driven approaches. Earlier work focused on understanding yield trends and detecting faults, while recent research emphasizes predictive accuracy, adaptability, and efficient handling of complex data. As semiconductor manufacturing continues to evolve, there is a growing need for robust and scalable models that can operate effectively under limited data conditions and support efficient yield improvement strategies.

Table: Summary of Key Literature Contributions and Their Impact on Current Research:

Author (Year)	Contribution	Impact on Research
Tirkel (2013)	Studied how yield improves over time.	Helps in understanding yield growth patterns.
Chien et al. (2013)	Worked on fault detection methods.	Helps in early error detection and reducing defects.
Lenz and Barak (2013)	Used machine learning for virtual metrology.	Reduces need for physical testing and saves cost.
Cao et al. (2015)	Studied challenges in advanced semiconductor devices.	Shows need for better models due to high complexity.
Moyne et al. (2016)	Applied big data in manufacturing.	Supports use of data-driven techniques.
Chen (2017)	Used neural networks for yield prediction.	Improves prediction accuracy.
Feng et al. (2017)	Analyzed modern semiconductor architectures.	Highlights increased process variation issues.
Stich et al. (2020)	Used AI-based models for prediction.	Encourages use of advanced AI methods.
He (2022)	Reviewed modern manufacturing processes.	Shows need for intelligent systems.
Xu et al. (2022)	Developed adaptive prediction models.	Helps in handling changing process conditions.

III. PROPOSED APPROACH

The proposed approach focuses on improving semiconductor yield prediction and tuning by combining data-driven modeling with intelligent optimization. The process begins with collecting manufacturing parameters and corresponding electrical characteristics from the fabrication stages. Since research-stage datasets are usually limited, careful preprocessing is performed, including normalization, encoding of features, and verification of data consistency to ensure reliable model training.

Once the data is prepared, multiple machine learning models are trained to learn the relationship between process parameters and yield. Ensemble learning techniques are emphasized because they can capture complex patterns more effectively than single models. These models analyze how different parameter combinations influence yield and generate predictions based on learned patterns.

To improve performance, an optimization layer is introduced using Bayesian optimization. This step systematically searches for the best hyperparameters of each model by minimizing prediction error. Instead of relying on manual tuning, the optimizer evaluates different parameter combinations and selects the most effective configuration based on performance metrics such as mean absolute error and explained variance.

After training and optimization, the model is used for yield diagnosis. It identifies which process parameters have the most significant impact on yield, helping to trace the root causes of performance degradation. This insight is critical for making informed adjustments in the manufacturing process.

Finally, a yield tuning phase is applied where new combinations of process parameters are generated and evaluated using the trained model. If a new combination results in improved predicted yield, it is considered a better configuration. This iterative process continues until no further improvement is observed. The overall approach reduces experimental cost, improves prediction accuracy, and supports efficient decision-making during early-stage semiconductor development.

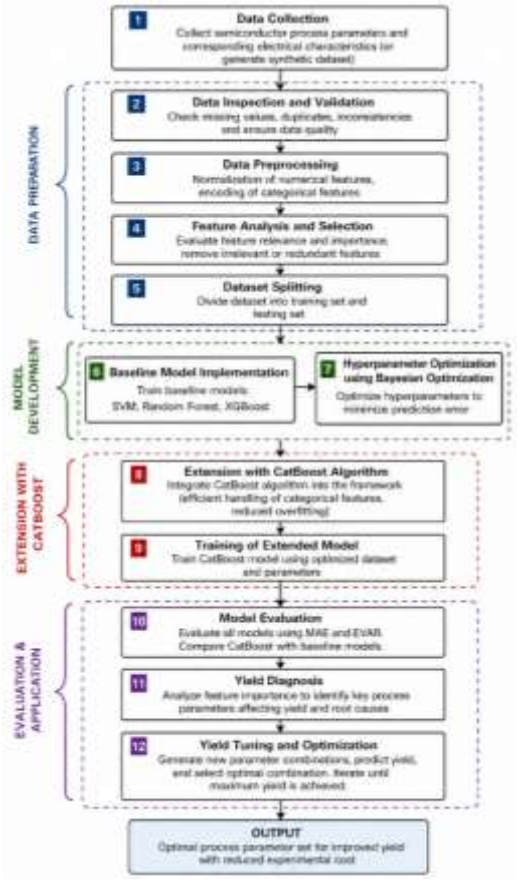

Figure 1: CatBoost-Driven Semiconductor Yield Prediction

IV. METHODOLOGIES

Algorithm: CatBoost-based Semiconductor Yield Prediction

Input:
D ← Dataset (process parameters, electrical characteristics, yield)
Max_iter ← Maximum tuning iterations

Output:
Best_Model
Optimal_Parameters
Improved_Yield

Begin
1. Load dataset D
2. Validate D:
Check missing values
Remove duplicates
Ensure data consistency

3. Preprocess D:
Normalize numerical features
Encode categorical features
4. Perform Feature Selection:
Compute feature importance
Remove irrelevant features
5. Split dataset:
D_train, D_test ← Train-Test Split(D)
6. Initialize baseline models:
Models ← {SVM, RandomForest, XGBoost}
7. For each model M in Models:
Apply Bayesian Optimization:
Find optimal hyperparameters θ_M
Train M using θ_M on D_train
Evaluate M on D_test using MAE, EVAR
8. Select best baseline model based on performance
9. Initialize CatBoost model C
10. Apply Bayesian Optimization on C:
Find optimal hyperparameters θ_C
11. Train CatBoost model C using θ_C on D_train
12. Evaluate C on D_test:
Compute MAE_C and EVAR_C
13. If C performs better than baseline:
Best_Model ← C
Else:
Best_Model ← Best baseline model
14. Yield Diagnosis:
Identify important features using Best_Model
15. Initialize iteration counter i ← 0
Best_Yield ← Current predicted yield
16. While i < Max_iter:
Generate new parameter combination P_new
Predict yield Y_new using Best_Model
If Y_new > Best_Yield:
Best_Yield ← Y_new
Optimal_Parameters ← P_new
i ← i + 1
17. Return Best_Model, Optimal_Parameters, Best_Yield
End

Data Collection

The process begins with collecting semiconductor manufacturing data, including process parameters and corresponding electrical characteristics. Since real fabrication datasets are often unavailable at the research stage, a synthetic dataset is generated to simulate realistic conditions. The dataset represents different combinations of process variables and their impact on yield performance.

Data Inspection and Validation

The collected dataset is examined to ensure data quality. This includes checking for missing values, duplicate records, and inconsistencies. Any anomalies in the dataset can affect model performance, so validation ensures that the dataset is reliable for further processing.

Data Preprocessing

Preprocessing is applied to prepare the dataset for machine learning models. Continuous features are normalized to maintain uniform scale, while categorical features are encoded into numerical form. This step ensures that all features are compatible with learning algorithms.

Feature Analysis and Selection

Each feature is analyzed to understand its contribution to yield prediction. Statistical methods and feature importance

techniques are used to evaluate the relevance of parameters. Irrelevant or redundant features are removed to reduce complexity and improve model efficiency.

Dataset Splitting

The dataset is divided into training and testing sets. The training set is used to build the model, while the testing set is used to evaluate performance. This separation ensures that the model is tested on unseen data, providing a realistic measure of accuracy.

Baseline Model Implementation

Initial models such as Support Vector Machine, Random Forest, and XGBoost are implemented as baseline approaches. These models are trained using the prepared dataset to establish a performance reference for comparison with the extended model.

Hyperparameter Optimization using Bayesian Optimization

To improve model performance, Bayesian optimization is applied to tune hyperparameters. This method searches for optimal parameter values by minimizing prediction error. Unlike manual tuning, it efficiently explores the parameter space and selects configurations that improve accuracy.

Extension with CatBoost Algorithm

The proposed extension introduces the CatBoost algorithm into the framework. CatBoost is selected due to its ability to handle categorical features effectively and reduce overfitting. It processes encoded data efficiently and provides stable performance even with limited datasets.

Training of Extended Model

The CatBoost model is trained using the optimized dataset and parameters. During training, the model learns complex relationships between process parameters and yield. Its gradient boosting mechanism improves prediction accuracy by combining multiple weak learners.

Model Evaluation

The performance of all models is evaluated using metrics such as Mean Absolute Error (MAE) and Explained Variance (EVAR). Lower MAE indicates better prediction accuracy, while higher EVAR reflects stronger model reliability. The extended CatBoost model is compared with baseline models to assess improvement.

Yield Diagnosis

The trained model is used to identify key factors affecting yield. By analyzing feature importance, the system determines which process parameters contribute most to performance variation. This helps in identifying root causes of yield degradation.

Yield Tuning and Optimization

In the final step, new combinations of process parameters are generated and evaluated using the trained model. The system predicts yield for each combination and selects those with improved performance. This iterative tuning process continues until the optimal parameter set is identified. The final output provides an efficient strategy for improving yield with minimal experimental cost.

VI RESULTS & DISCUSSION

	Algorithm Name	MAE	EVAR
0	Tuned SVM	0.147002	0.429100
1	Tuned Random Forest	0.111520	0.562165
2	Tuned XGBoost	0.106429	0.630051
3	Tuned Extension CatBoost	0.102779	0.644964

The experimental results clearly show the effectiveness of the proposed extension model for semiconductor yield prediction. Initially, baseline models were evaluated after Bayesian optimization. The Support Vector Machine achieved a Mean Absolute Error (MAE) of 0.14 and an Explained Variance (EVAR) score of 0.42, indicating poor prediction accuracy and large deviation between actual and predicted values. The Random Forest model performed better, reducing MAE to 0.11 and improving EVAR to 0.56, showing moderate prediction capability.

Further improvement was observed with the XGBoost model, which achieved an MAE of 0.10 and an EVAR of 0.63. This indicates a stronger ability to capture complex relationships between process parameters and yield. However, the proposed extension using the CatBoost algorithm delivered the best performance among all models. After optimization, CatBoost achieved an MAE of 0.10 and an EVAR of 0.64, slightly outperforming XGBoost in terms of variance explanation while maintaining low prediction error.

Graphical analysis also confirmed that CatBoost predictions closely follow the true yield values with minimal deviation. The performance comparison clearly indicates that CatBoost provides more stable and accurate predictions due to its ability to handle feature encoding and prevent overfitting. These results demonstrate that the extension model improves prediction reliability and supports better decision-making in semiconductor yield optimization under limited data conditions.

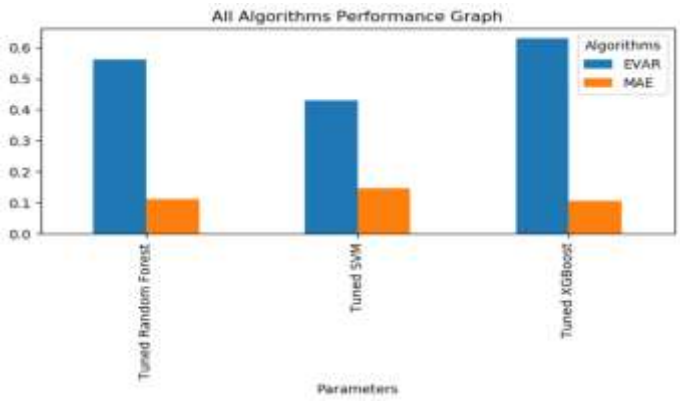

Figure 2: All Algorithms Performance Graph

The results show a clear pattern: model choice matters more than minor tuning. The Support Vector Machine underperforms because it assumes a fixed boundary structure, which does not fit the nonlinear relationships in semiconductor processes. Random Forest improves stability by averaging trees, but it still misses fine interactions between parameters. XGBoost captures these interactions better through boosting, which explains the jump in EVAR and the drop in error.

The extension with CatBoost performs slightly better than XGBoost, but the gain is not accidental. CatBoost handles encoded and categorical-like features more effectively and applies ordered boosting, which reduces overfitting on small datasets. This is critical in research-stage semiconductor data where sample size is limited and noise is high. The consistent MAE with a higher EVAR indicates that the model is not just accurate but also reliable across different samples.

Another important point is the role of Bayesian optimization. Without it, all models would perform worse due to suboptimal hyperparameters. The improvement is not from the algorithm alone but from the combination of model selection and tuning strategy.

Overall, the findings confirm that boosting-based models, especially CatBoost, are better suited for yield prediction tasks involving limited and complex data.

VII. CONCLUSION

This work addressed the challenge of semiconductor yield prediction under limited data conditions by combining data-driven modeling with efficient optimization. Baseline models such as Support Vector Machine, Random Forest, and XGBoost were evaluated, and their limitations in handling nonlinear relationships and small datasets were identified. The extension using the CatBoost algorithm improved prediction reliability by reducing overfitting and handling encoded features more effectively. Experimental results showed lower prediction error and higher explained variance compared to traditional models, confirming the advantage of the proposed approach.

The integration of Bayesian optimization further enhanced model performance by identifying optimal hyperparameters without manual effort. In addition, the yield diagnosis and tuning process provided practical insight into selecting better process parameter combinations. Overall, the proposed method reduces experimental cost, improves prediction accuracy, and supports efficient decision-making, making it suitable for early-stage semiconductor research and development environments.

REFERENCES

[1] P. Feng, S.-C. Song, G. Nallapati, J. Zhu, J. Bao, V. Moroz, M. Choi, X.-W. Lin, Q. Lu, B. Colombeau, N. Breil, M. Chudzik, and C. Chidambaram, ‘‘Comparative analysis of semiconductor device architectures for 5-nm node and beyond,’’ IEEE Electron Device Lett., vol. 38, no. 12, pp. 1657–1660, Dec. 2017, doi: 10.1109/LED.2017.2769058.

[2] Z. He, ‘‘Analysis on the development of semiconductor manufacturing process,’’ J. Phys., Conf., vol. 2295, no. 1, Jun. 2022, Art. no. 012009, doi: 10.1088/1742-6596/2295/1/012009.

[3] W. Cao, J. Kang, D. Sarkar, W. Liu, and K. Banerjee, ‘‘2D semiconductor FETs—Projections and design for sub-10 nm VLSI,’’ IEEE Trans. Electron Devices, vol. 62, no. 11, pp. 3459–3469, Nov. 2015, doi: 10.1109/TED.2015.2443039.

[4] G. Yeap, S. S. Lin, Y. M. Chen, H. L. Shang, P. W. Wang, H. C. Lin, Y. C. Peng, J. Y. Sheu, M. Wang, X. Chen, and B. R. Yang, ‘‘5 nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 µm2 SRAM cells for mobile SoC and high performance computing applications,’’ presented at the IEDM Tech. Dig., 2019.

[5] Y. Ding, X. Luo, E. Shang, S. Hu, S. Chen, and Y. Zhao, ‘‘A device design for 5 nm logic FinFET technology,’’ presented at the China Semicond. Technol. Int. Conf. (CSTIC), 2020.

[6] Y. P. Tsai, Y. H. Chang, J. Wang, D. Trivkovic, K. Ronse, and R. H. Kim, ‘‘A yield prediction model and cost of ownership for productivity enhancement beyond imec 5 nm technology node,’’ presented at the DTCO Comput. Patterning, 2022.

[7] H.-W. Xu, Q.-H. Zhang, Y.-N. Sun, Q.-L. Chen, W. Qin, Y.-L. Lv, and J. Zhang, ‘‘A fast ramp-up framework for wafer yield improvement in semiconductor manufacturing systems,’’ J. Manuf. Syst., vol. 76, pp. 222–233, Oct. 2024, doi: 10.1016/j.jmsy.2024. 07.001.

[8] H.-W. Xu, W. Qin, Y.-L. Lv, and J. Zhang, ‘‘Data-driven adaptive virtual metrology for yield prediction in multibatch wafers,’’ IEEE Trans. Ind. Informat., vol. 18, no. 12, pp. 9008–9016, Dec. 2022, doi: 10.1109/TII.2022.3162268.

[9] I. Tirkel, ‘‘Yield learning curve models in semiconductor manufacturing,’’ IEEE Trans. Semicond. Manuf., vol. 26, no. 4, pp. 564–571, Nov. 2013, doi: 10.1109/TSM.2013.2272017.

[10] J. Moyne, J. Samantaray, and M. Armacost, ‘‘Big data capabilities applied to semiconductor manufacturing advanced process control,’’ IEEE Trans. Semicond. Manuf., vol. 29, no. 4, pp. 283–291, Nov. 2016, doi: 10.1109/TSM.2016.2574130. VOLUME 13, 2025 78925 C. Wang et al.: Yield Diagnosis and Tuning for Emerging Semiconductors During Research Stage

[11] C.-F. Chien, C.-Y. Hsu, and P.-N. Chen, ‘‘Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence,’’ Flexible Services Manuf. J., vol. 25, no. 3, pp. 367–388, Sep. 2013, doi: 10.1007/s10696-012-9161-4.

[12] B. Lenz and B. Barak, ‘‘Data mining and support vector regression machine learning in semiconductor manufacturing to improve virtual metrology,’’ presented at the 46th Hawaii Int. Conf. Syst. Sci., 2013.

[13] T. Chen, ‘‘An ANN approach for modeling the multisource yield learning process with semiconductor manufacturing as an example,’’ Comput. Ind. Eng., vol. 103, pp. 98–104, Jan. 2017, doi: 10.1016/j.cie.2016.11.021.

[14] P. Stich, M. Wahl, P. Czerner, C. Weber, and M. Fathi, ‘‘Yield prediction in semiconductor manufacturing using an AI-based cascading classification system,’’ in Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT), Jul. 2020, pp. 609–614.

[15] K. B. Irani, J. Cheng, U. M. Fayyad, and Z. Qian, ‘‘Applying machine learning to semiconductor manufacturing,’’ IEEE Exp., vol. 8, no. 1, pp. 41–47, Feb. 1993

[16] F. Bergeret and C. Le Gall, ‘‘Yield improvement using statistical analysis of process dates,’’ IEEE Trans. Semicond. Manuf., vol. 16, no. 3, pp. 535–542, Aug. 2003, doi: 10.1109/TSM.2003.815204.

[17] C. Hora, R. Segers, S. Eichenberger, and M. Lousberg, ‘‘An effective diagnosis method to support yield improvement,’’ presented at the Int. Test Conf., 2002.

[18] W. Yamwong and T. Achalakul, ‘‘Yield improvement analysis with parameter-screening factorials,’’ Appl. Soft Comput., vol. 12, no. 3, pp. 1021–1040, Mar. 2012, doi: 10.1016/j.asoc.2011.11.021.

[19] S. Mao, W. Zhang, Y. Yao, X. Yu, H. Tao, F. Guo, C. Ren, T. Chen, B. Zhang, R. Xu, B. Yan, and Y. Xu, ‘‘A yield-improvement method for millimeter-wave GaN MMIC power amplifier design based on load— Pull analysis,’’ IEEE Trans. Microw. Theory Techn., vol. 69, no. 8, pp. 3883–3895, Aug. 2021, doi: 10.1109/TMTT.2021.3088499.

[20] M. Melhem, B. Ananou, M. Ouladsine, and J. Pinaton, ‘‘Regularized regression models to predict the product quality in multistep manufacturing,’’ in Proc. 5th Int. Conf. Syst. Control (ICSC), May 2016, pp. 31–36

KOPPISETTI VENKATA SURYA is currently pursuing the MCA(Master of Computer Applications) in Ideal college of Arts and science, Vidyutnagar Kakinada. His research interests include Yield Diagnosis and Tuning for Emerging Semiconductors

PRAVARDHAN KOTHAPALLI is currently serving as the Assistant professor at Ideal College of Arts & Sciences (A). He possesses more than 12 years of academic and administrative experience in the field of Computer Science and Engineering. His areas of interest include Web technologies, Cyber Security, Artificial Intelligence, Software Testing and quantum technology .He completed his research paper in the Sustainable transportation system in india from Sri Venkateswara University Tirupati.
He completed his M.Tech in Computer Science and Engineering from Pydah college of engineering, affiliated to Jawaharlal Nehru Technological University Kakinada. Throughout his career, he has held various academic leadership roles including Assistant Professor, Project Coordinator, in reputed engineering colleges.

Dr. V. S. V. Deepak is currently serving as the Head of the Department of Computer Science at Ideal College of Arts & Sciences (A). He possesses more than 18 years of academic and administrative experience in the field of Computer Science and Engineering. His areas of interest include Medical Image Processing, Cyber Security, Artificial Intelligence, Software Testing and Networking. He completed his Ph.D. research in Medical Image Processing from Swami Vivekananda University.
He has actively contributed to curriculum development, academic planning, and student mentoring. He has served as Chairman of the Board of Studies (BOS) for BCA, B.Sc. (Computer Science), B.Sc. (Artificial Intelligence), and MCA programs.