Threshold Selection for Classification Models in Prognostics

Rohit Deo, Swarali Desai, Subhalakshmi Behera, Chetan Pulate, Aman Yadav and Nilesh Powar

Cummins Technologies India Pvt. Ltd., Pune, MH, 411045, India

Rohit.deo@cummins.com
Swarali.desai@cummins.com
Subhalakshmi.behera@cummins.com
Chetan.pulate@cummins.com
Aman.yadav@cummins.com
Nilesh.powar@cummins.com

ABSTRACT

In this study, we evaluate the performance of a prognostic classification model for NOX sensors in diesel engines over one month by comparing its predictions against actual outcomes. We then construct a validation dataset to assess the model's performance. By analyzing instances where the model's predictions were incorrect, we determine new threshold values that could potentially reduce errors for each false positive (FP) and false negative (FN). Subsequently, we create a dataset where the threshold varies for each observation and train a regression model with the modified threshold as the target variable. Our findings indicate that incorporating this approach, where the model's performance is iteratively refined using the validation dataset, leads to a reduction in both false positives and false negatives.

Keywords – True Negative (TN), True Positive (TP), False Negative (FN), False Positive (FP), Receiver Operating Characteristic (ROC), Area Under ROC Curve - (AUC)

1. INTRODUCTION

Cummins Inc. is a global corporation that designs, manufactures, and distributes engines, filtration, and power generation products. Cummins Inc. is headquartered in Columbus, Indiana, and has a history dating back to 1919. The company serves customers in more than 190 countries and territories, with a focus on innovation and sustainability in its products and operations. Prognostics plays a crucial role for Cummins in the context of diesel engines by enabling predictive maintenance. By analyzing the condition and performance of diesel engines using data from sensors and other sources, prognostics can help Cummins predict when maintenance or repairs will be needed.

This predictive approach allows Cummins to schedule maintenance in advance, minimizing downtime and reducing the risk of unexpected failures. Overall, prognostics help Cummins optimize the performance, reliability, and longevity of their diesel engines.

Diesel engines, the preferred power source for commercial vehicles like trucks and buses, produce harmful NO and NO2 emissions due to high combustion temperatures. Mckinley, Somwanshi, Bhave, and Verma, (2020) in their study showed that, to meet stringent emission standards, after-treatment systems such as selective catalytic reduction (SCR) are used, which can reduce emissions by factors of 10 to 20. SCR involves injecting a Diesel Exhaust Fluid (DEF) into the exhaust to produce ammonia (NH3), which then reacts with NOX to form harmless nitrogen (N2). NOX sensors are crucial in this process, measuring conversion efficiency and guiding the injection rate of DEF. Errors in these sensors can lead to either excessive ammonia or NOX emissions, impacting air quality and health. Regulatory agencies require continuous monitoring of these sensors and their operation to ensure compliance.

Prognostics aims to suggest changing out a NOX sensor before it fails. Since the replacement NOX Sensor can be planned for a convenient time rather than dealing with the discomfort of an unexpected breakdown, the customer will ideally experience less downtime. To determine whether the NOX sensor will fail, the present prognostics methodology uses a classification model with a predetermined threshold which is set using AUC ROC curve analysis. Even with an appropriate threshold, there may be instances where the model's predictions are not entirely accurate, which could potentially lead to increased downtime and maintenance costs for the customer. Dynamic thresholding model is a field of research that focuses on developing efficient methods for altering decision thresholds in predictive models over time and over different units.

The goal is to reduce the cost of false positives and negatives (after observing the performance of the prognostic classification model) by accounting for changes in the underlying data distributions that can arise because of changes in the environment, warranty status, or other external factors. In this study, we investigated a dynamic thresholding strategy and measured its performance by evaluating the incremental financial impact on customers and businesses. We also suggested a method for constructing a target label based on prognostics likelihood and validation data by computing the optimum thresholds. Our results demonstrate the importance of dynamic thresholding in maintaining the accuracy and robustness of predictive models and highlight the potential for further improvements through continued research in this area.

In summary, the paper explains how to dynamically change the threshold to improve the performance of a classification model after observing its performance for some time.

The rest of the paper is organized as follows. Section 2 gives a literature survey. Section 3 elaborates proposed Dynamic Thresholding Model (DTM) followed by results and discussions in section 4. Section 5 gives conclusions and future scope.

2. Literature Survey

In this literature review, we will explore some of the key research papers that use ROC and AUC, dynamic thresholding, and cost analysis to determine thresholds.

Threshold selection is a crucial step in binary classification models as it determines the balance between the trade-off of precision and recall. Receiver operating characteristic (ROC) curves and area under the curve (AUC) are commonly used metrics to evaluate the performance of binary classification models and determine the optimal threshold value. The concept of the ROC (Receiver Operating Characteristic) curve, from which AUC ROC is derived, originated in electrical engineering and signal detection theory. It was initially used to analyze the performance of radar systems during World War II. The ROC curve was later adopted in medicine to evaluate diagnostic tests' performance. In machine learning, the ROC curve is used to assess the performance of binary classification models. The AUC ROC is a numerical measure derived from the ROC curve and provides a single value to quantify the overall performance of a classifier.

The ROC curve and AUC ROC can help in deciding the threshold for a binary classification model by providing insights into the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different threshold values. In their classic paper, Bradley (1997) emphasizes that AUC ROC provides a comprehensive measure of a classifier's performance across all possible thresholds, making it particularly useful for assessing the overall discriminatory ability of a model. The paper highlights that AUC ROC can be instrumental in threshold selection by illustrating the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) at different threshold values. This insight enables practitioners to choose an optimal threshold based on the specific needs of the classification problem, balancing the costs associated with false positives and false negatives. In addition, Bradley discusses how AUC ROC can help in selecting a threshold that best suits the application's requirements. By analysing the ROC curve, which plots the true positive rate against the false positive rate at various thresholds, practitioners can visualize the classifier's performance and make informed decisions about threshold selection. This capability is particularly valuable in scenarios where the cost of false positives and false negatives differs, as it allows for the customization of the classifier's behaviour to meet specific needs.

Alotaibi and Flach (2021) introduce a novel approach to extend the traditional AUC metric to incorporate misclassification costs, addressing limitations in existing settings. By treating costs as sampled data, the proposed method employs the Weighted AUC (WAUC) metric and a novel estimator to approximate it, enabling a more accurate representation of model performance in complex cost-sensitive scenarios. The approach establishes a correspondence between WAUC and the cost function using threshold weighting and presents a bilevel optimization formulation to couple them. This formulation ensures that WAUC can be optimized at the optimal threshold value based on the real-world cost distribution. A stochastic algorithm is proposed for optimizing this formulation, demonstrating convergence rates comparable to standard SGD. Experimental results validate the effectiveness of the method in extending AUC to cost-sensitive scenarios, highlighting its significant performance improvements.

Yang, Yu, Wang, Quddus and Xue (2018) introduced the thresholding methods: fixed, rate-driven, optimal, RCut, MCut, and two novel ones: score-driven and global optimal are introduced. Score-driven thresholds can be adjusted globally or per label, offering flexibility. It investigates selecting a single global threshold or multiple thresholds. Using real-world datasets, the study conducts an empirical review, finding that the global and label-wise score-driven methods excel. Tuning a global threshold with respect to per-label cost is not significantly worse than using a separate threshold per label. Some traditional approaches, like the label-wise rate-driven method, may not suit highly imbalanced multi-label data. The study recommends using score-driven thresholds, globally or per label, for superior performance. It calls for further research on misclassification costs, loss, and threshold choice in multi-label classification, particularly when costs vary across labels.

In their study, Johnson and Khoshgoftaar (2019) showed that class imbalance is a common issue in machine learning, addressed through algorithm-level, data-level, and hybrid methods. While extensively studied in traditional algorithms, its application to deep neural networks (DNNs) is limited. This paper fills this gap by studying thresholding in DNNs using a Big Data Medicare fraud dataset. Employing random oversampling (ROS), random under-sampling (RUS), and a hybrid ROS-RUS, 15 training distributions with varying imbalance levels are created. Optimal classification thresholds are identified for each distribution on random validation sets, outperforming default thresholds. They further showed that, statistical analysis reveals a strong linear relationship between minority class size and optimal threshold, highlighting the importance of thresholding in DNNs for imbalanced data.

The properties of the F1 performance metric in multilabel classification, particularly regarding optimal decision-making thresholds. In this study, Lipton, Elkan, and Narayanswamy (2014) discuss how the best achievable F1 score is linked to the optimal threshold and highlights the impact of classifier behaviour in uninformative scenarios. For instance, in such scenarios, predicting all instances as positive maximizes the expectation of F1, which is beneficial for some metrics but problematic for others, like macro F1 in the presence of rare labels. The study also reveals that micro F1, on the other hand, maximizes the expected score by predicting all examples as negative in similar scenarios. This insight is especially valuable in settings with numerous labels. Additionally, the study suggests that micro F1 may wash out performance on rare labels. The findings underscore the importance of carefully selecting and understanding performance metrics, especially when choosing a single metric to optimize in scenarios involving competing systems, as this choice can significantly impact optimal thresholding behavior.

For a different application, Hancock, Johnson and Khoshgoftaar (2022) investigate the impact of the TPR ≥ TNR constraint on threshold values in classification tasks. The constraint favors lower thresholds, closer to the prior probability of the positive class, leading to reasonable trade-offs in classification rates. The default decision threshold of 0.5 is found unsuitable for the imbalanced Kaggle Credit Card Fraud Detection Dataset, yielding low TPR and FNR scores. It is noted that this default threshold is much larger than the prior probability of the positive class in imbalanced data. No single metric provides a comprehensive view of classifier performance, with thresholds closer to the positive class prior probability generally yielding better performance across multiple metrics. Each threshold selection technique offers trade-offs between positive and negative class performance. Starting with the positive class prior probability as a benchmark, thresholds can be adjusted to balance TPR and TNR scores. For specific performance goals, the optimal threshold can be estimated using the training dataset, considering user-defined performance metrics and constraints. The choice of performance metrics and constraints for threshold optimization significantly impacts test performance, highlighting the importance of careful selection based on the classification task's requirements and goals.

Going back in time, Chen, Tsai, Moon, Ahn, Young, and Chen (2006) explore the impact of decision thresholds on sensitivity, specificity, and concordance in four classification methods: logistic regression, classification tree, Fisher's linear discriminant analysis, and weighted k-nearest neighbour. While standard classification algorithms aim to maximize correct predictions (concordance), this may not be suitable for all applications. Some applications prioritize high sensitivity (e.g., clinical diagnostics), while others prioritize high specificity (e.g., epidemiology screening studies). The study examines the use of decision threshold adjustment to enhance sensitivity or specificity under specific conditions. Through Monte Carlo simulations, the study shows that increasing the decision threshold leads to decreased sensitivity and increased specificity, with concordance values remaining stable within an interval around the maximum concordance. Optimal decision thresholds can be identified within this interval to meet specified sensitivity and specificity requirements. The study analyzes three example datasets to illustrate these findings.

Two variants are introduced: a novel neural network-based thresholding method called ThresNets for improving multi-label predictions from class scores obtained from external scorers as shown by Shao & Huiyang et al. (2024) ThresNets are designed to scale linearly with the number of labels and can be trained offline after the scorer training is completed. One variant incorporates classic CS/CSS thresholds into the neural model, serving as a form of transfer learning between heterogeneous models. Our method is particularly suitable for medium-sized multi-label classification (MLC) tasks where informative label score dependencies can be found, and the ground truth of label assignments is reliable. Experimental results on artificially created scores demonstrate the effectiveness of ThresNets, especially when the scoring phase allows for improvements. ThresNets outperformed popular nearest neighbor-based classifiers in recovering from scoring errors. Empirical evaluation on real datasets shows that ThresNets perform better than classical methods according to various metrics, especially when used in a hybrid approach. Despite its advantages, ThresNets face challenges such as the risk of overfitting and the difficulty of training with long-tail labels. Future work will focus on leveraging external knowledge of class structure to improve ThresNets' performance further.

In their classic Real-time crash prediction, Draszawka, Karol and Szymanski (2023) show how important the Threshold selection is, by determining the cut-off point for the posterior probability used to separate potential crash warnings. Current research lacks methods for effectively determining an optimal threshold, often resorting to subjective approaches. This study proposes a theoretical method using the mixed logit model to develop crash risk evaluation models. The minimum cross-entropy method outperforms other threshold selection methods, providing a reliable and automatic approach for identifying optimal thresholds in crash prediction.

In conclusion, this literature review has explored various aspects of threshold selection in machine learning, particularly in the context of class imbalance and multi-label classification. The reviewed papers have highlighted the importance of selecting an appropriate threshold for optimizing model performance and addressing specific challenges in different applications.

Several studies have proposed novel thresholding methods, such as ThresNets for multi-label classification and the use of the Area Under the ROC Curve (AUC ROC) metric for threshold selection. These methods have shown promising results in improving classification performance, especially in scenarios with imbalanced data and complex cost-sensitive considerations.

However, challenges such as overfitting and training with long-tail labels remain, suggesting the need for further research. Future studies could focus on leveraging external knowledge, such as class structure information, to enhance thresholding methods. Additionally, exploring the application of thresholding in emerging areas like real-time crash prediction and credit fraud detection could lead to valuable insights and advancements in the field.

3. Proposed Methodology

The prognostics model may make a future failure prediction for each component. We can run the prognostics model, determine the component's remaining useful life (RUL), and take preventive action. The model predictions for this classification model are based on the threshold and the likelihood for each data row. The threshold is often set using AUC-ROC curves, and that threshold is chosen for which the area under the curve for the TPR vs. FPR curve is highest. Our current prognostics algorithm alerts the client 90 days before the component's likely failure.

The dataset has two variations, pre-verified engine validation data, and post-verified engine validation data. Both have all engine status parameters, but they differ on when the actual status of the engine was captured. In the pre-verified engine validation data we predict engine health using a static threshold to classify an engine as healthy or faulty after we have the data about the actual status of the engine. In post-verified engine validation data, we predict before we have any ground truth for engine health using a static threshold to classify an engine as healthy or faulty, the ground truth is received later. The pre-verified engine validation data has been discarded due to data quality issues. The post-verified engine validation data has features such as the likelihood of an engine failure, the RUL of the engine, the classification threshold that was applied, the prediction by the predictive maintenance model, the actual status of the engine, and all other parameters of engine health.

Based on validation data in the suggested approach, we can choose the threshold dynamically. The validation data, with which, once the prognostics model has been run, we validate its performance during the following 90 days. Following the 90 days, we learn whether the failed event occurred. The confusion matrix, which we use to interpret the TPs, FPs, TNs, FNs, etc., is provided by the validation data. Our suggested methods can further reduce the FPs and FNs.

The order of events is depicted in the block diagram. The dynamic thresholding model is constructed utilizing features like odometer running, engine run duration, the likelihood of each row, the component's warranty status, and failure type (first or repeated failure), as opposed to setting a

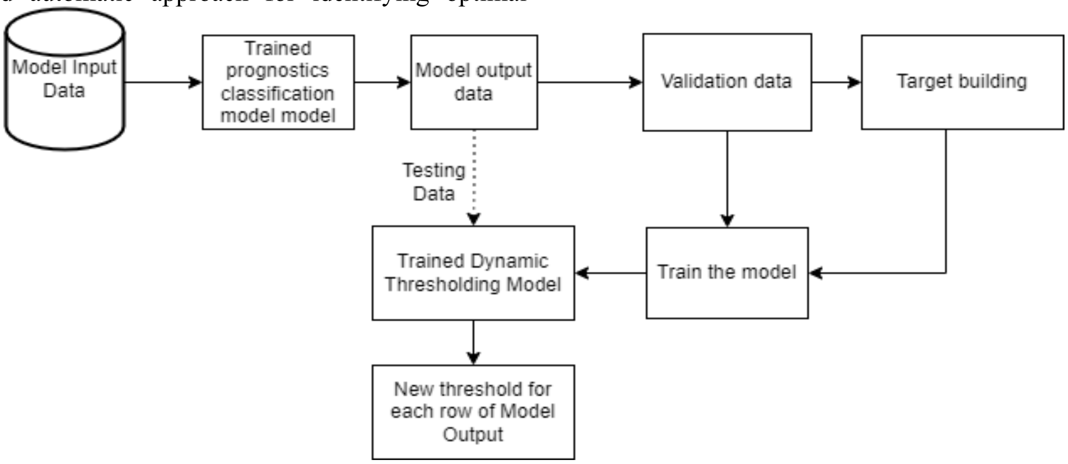

Figure 1. Block diagram of Dynamic Thresholding Model

threshold for the entire dataset. A regression model is the dynamic thresholding model. Concerning the validation data, the following algorithm is used to construct the regression model's target.

1. *For a TP and TN, there is no change in the threshold, fixed by the AUC-ROC curve.*
2. *E.g. - likelihood = 0.76, the fixed threshold is 0.8, which comes out to FN in the validation data, then, the new threshold is calculated as –*

$$Threshold_{New} = Threshold_{old} + |Threshold_{old} - Likelihood| - correction\_factor^{\#}$$

3. *Similarly, if likelihood = 0.86, the fixed threshold is 0.8, which comes out to FP in the validation data, then, the new threshold is calculated as –*

$$Threshold_{New} = Threshold_{old} + |Threshold_{old} - Likelihood| + correction\_factor^{\#}$$

4. *Repeat the procedure for each TP, FP, TN, and FN.*
5. # - The value can range between 0 to 1. This can be finalized after repeated training and testing of the model for different correction factors.

The target label, a continuous variable with a distinct threshold for each row of the output, is trained with the regressor model after we create the dataset containing the features as described before. After the model has been trained on the validation data, the DTM model that predicts dynamic threshold is given model out data with the same attributes as the validation data.

Extreme Gradient Boosting (XGBoost), a powerful machine learning method that we have used for this regression problem, is capable of handling complex non-linear interactions between features and targets as well as handling missing values and outliers.

The performance of the Xgb regressor used in DTM, the effect of changing the threshold, and the financial impact of this approach for each row are discussed in detail in the results and discussions.

4. Results and Discussions

The results (in Table 1) after applying Dynamic Threshold Modeling (DTM) show a notable improvement in various metrics. We have compared our results with traditional AUC ROC curves, as shown in column 2 of Table 1. We obtained the optimised threshold to be zero when we experimented with weighted AUC (WAUC) curves. This means that the model was recommending to replace every NOX sensor with even the smallest likelihood for failure. This is not a pragmatic solution. Also, The score-driven global thresholds proved to give no different results than the traditional AUC. Because of data availability constraints per the rest of the methodologies, we couldn't compare our methodology with them.

As compared to AUC, the True Positive Rate (TPR) has increased from 0.069 to 0.156, indicating that the model is better at correctly identifying positive instances. The Precision has also improved significantly, rising from 0.041 to 0.093, indicating a reduction in false positives. This improvement is further reflected in the F1 Score, which has increased from 0.074 to 0.14. Despite these improvements, the model still exhibits a relatively high False Positive Rate (FPR), albeit reduced from 0.397 to 0.373.

Table 1 shows the results for the EONOX sensor of a popular engine series. We observe that a reduction in FPs saves unnecessary repair of the engines and a reduction in FNs saves downtime cost of the engines which saves $0.3M for our customers and $13k for our company. In total, we save $313k. The methodology is highly scalable.

Overall, applying DTM has greatly enhanced the model's performance, particularly in correctly identifying positive instances and reducing false positives.

Table 1. Comparison of TPs, TNs, FPs and FNs before and after the use of DTM

	Using (AUC-ROC)	After DTM
TP	167	377
FP	3905	3670
FN	2246	2036
TN	5923	6158
TPR	0.069	0.156
FPR(or Recall)	0.397	0.373
Precision	0.041	0.093
F1 Score	0.074	0.14

5. Conclusions and Future Scope

The application of Dynamic Threshold Modeling (DTM) in our classification model has resulted in significant enhancements across key performance metrics, including accuracy, precision, and recall. By increasing the number of True Positives (TPs) and True Negatives (TNs) while decreasing False Positives (FPs) and False Negatives (FNs), the DTM has improved the model's overall effectiveness. We have implemented the DTM on one sensor from a specific family of engines used in a particular application. However, there is potential to expand this approach to multiple sensors across various engine families and to integrate it with other classification models. This scalability could save substantial costs by minimizing unnecessary repairs and downtime. We estimate that such an expansion could result in savings amounting to millions of dollars.

References

Mckinley, T., Somwanshi, M., Bhave, D. and Verma, S. 2020. *Identifying NOx Sensor Failure for Predictive Maintenance of Diesel Engines using Explainable AI*. PHM Society European Conference. 5, 1 (Jul. 2020), 11.

Andrew P. Bradley, *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition, Volume 30, Issue 7, 1997, Pages 1145-1159, ISSN 0031-3203.

Reem Alotaibi, Peter Flach, *Multi-label thresholding for cost-sensitive classification*, Neurocomputing, Volume 436, 2021, Pages 232-247, ISSN 0925-2312.

Kui Yang, Rongjie Yu, Xuesong Wang, Mohammed Quddus, Lifang Xue, *How to determine an optimal threshold to classify real-time crash-prone traffic conditions?, Accident Analysis & Prevention*, Volume 117, 2018, Pages 250-261, ISSN 0001-4575.

J. M. Johnson and T. M. Khoshgoftaar, "Deep Learning and Thresholding with Class-Imbalanced Big Data," 2019, *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA, 2019, pp. 755-762.

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). *Thresholding classifiers to maximize F1 score*. arXiv preprint arXiv:1402.1892.

J. Hancock, J. M. Johnson and T. M. Khoshgoftaar, "A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data," 2022, *IEEE 8th International Conference on Collaboration and Internet Computing (CIC), Atlanta*, GA, USA, 2022, pp. 135-142.

Chen, J. J., Tsai, C. A., Moon, H., Ahn, H., Young, J. J., & Chen, C. H. (2006). *Decision threshold adjustment in class prediction*. SAR and QSAR in Environmental Research, 17(3), 337–352.

Shao, Huiyang, et al. "Weighted roc curve in cost space: Extending auc to cost-sensitive learning." *Advances in Neural Information Processing Systems 36* (2024).

Draszawka, Karol, and Julian Szymański. "From Scores to Predictions in Multi-Label Classification: Neural Thresholding Strategies." *Applied Sciences 13.13 (2023): 7591.*

Biographies

Rohit Deo received his Bachelor's degree in Electronics and Telecommunication from WIT, Solapur in 2011 and his Master's degree in Signal Processing from the University of Pune. He has eight years of professional experience. Rohit began his career as an Assistant Professor at Modern College of Engineering, Pune from July 2014 to May 2018. He then transitioned to industry, working as a Data Scientist at Montran Corporation, Mumbai from July 2018 to December 2021. Since December 2021, he has been a Senior Data Scientist at Cummins Technologies, Pune. His current and previous research interests include predictive analysis, pattern classification, machine learning, and deep learning for complex business problems. Rohit is highly skilled in signal processing and applied mathematics. He is a member of various professional societies.

Swarali Desai is a Data Science and Digital Engineering professional at Cummins. She holds a bachelor's degree in Electronics and Telecommunications Engineering from the University of Mumbai, India, and a Master of Science in Data Science from the University of Washington. With extensive experience in data analysis, machine learning, and business intelligence, Swarali has led several impactful projects in predictive maintenance, engine failure prediction, and supply chain optimization. Her recent work involves applying advanced data analytics techniques to enhance the performance and reliability of Cummins' engine systems, thereby contributing significantly to operational efficiency and customer satisfaction.

Subhalakshmi Behra holds a B.Tech in Mechanical Engineering from NIT Rourkela (2008) and an MBA in Business Management from XLRI Jamshedpur (2010). With over 14 years of industry experience, she currently serves as the Analytics Manager for the Analytics and Artificial Intelligence team at Cummins Inc. Her expertise spans advanced analytics, machine learning, digital product development, supply chain management, and project management. Subha is a Certified Six Sigma Green Belt and has numerous awards and honors for her contributions to the field.

Chetan Pulate, holds a Bachelor's degree in Electronics and Telecommunication from Smt. Kashibai Navale College of Engineering. With over 6 years of experience at Cummins Inc. as a Data Engineer, Chetan specializes in developing distributed, scalable, and reliable data pipelines. His expertise includes Apache Spark, Hive, Sqoop, Oozie, MapReduce, and various Hadoop stacks, along with working in Microsoft Azure. Chetan is skilled in feature engineering and deploying data science models at scale. He has also worked at Cognizant and Capgemini, enhancing his skills in big data and data engineering.

Aman Yadav holds a Bachelor of Technology degree in Computer Science from the University Institute of

Engineering and Technology, Kanpur (2021), and a Master of Technology degree in Artificial Intelligence from the Defense Institute of Advanced Technology (DIAT), Pune (2023). He is a Data Scientist at Cummins India, where he has been working since August 2023, after completing an internship as an M.Tech AI intern at the same company. Aman has experience in developing AI models, including an ASR model for Indian accents, gained during his internship at ProxMaq. His skills include machine learning, deep learning, NLP, and working with technologies like Apache Spark, Scikit-Learn, and Microsoft Azure. Aman is passionate about advancing AI and robotics, and holds several certifications in graph data science from Neo4j.

Nilesh Powar is the Advanced Analytics Director at Cummins. He holds bachelor in electronics engineering from University of Bombay, India, M.S in Computer Engineering from Wright State University and a doctoral degree in Electrical and Computer Engineering from University of Dayton, Ohio. He has over 20+ years' experience in field of image processing, machine learning, statistical pattern recognition and system integration. He had worked in the US as Distinguished Research Scientist for University of Dayton Research Institute, Dayton, OH, USA. Recent efforts involve data analytics for die casting, predictive analysis for supply chain management and video summarization using deep learning.