IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 30 March 2024, accepted 4 June 2024, date of publication 8 July 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3425226

RESEARCH ARTICLE

Demystifying Defects: Federated Learning and Explainable AI for Semiconductor Fault Detection

TANISH PATEL[1], (Student Member, IEEE), RAMALINGAM MURUGAN[2], GOKUL YENDURI[3], RUTVIJ H. JHAVERI[1], (Senior Member, IEEE), HICHEM SNOUSSI[4], AND TAREK GABER[5], (Member, IEEE)

[1]Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar 382007, India
[2]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
[3]School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India
[4]LIST3N Research Unit, Université de Technologie de Troyes, 10300 Troyes, France
[5]School of Science, Engineering & Environment, University of Salford, M5 4WT Manchester, U.K.

Corresponding authors: Rutvij H. Jhaveri (rutvij.jhaveri@sot.pdpu.ac.in) and Tarek Gaber (t.m.a.gaber@salford.ac.uk)

This work was supported by the Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University.

ABSTRACT Semiconductor manufacturing, a critical driver of modern technology, involves intricate processes for fabricating integrated circuits on materials like silicon. This industry's pivotal role spans various applications, from smartphones to computers, emphasizing the importance of fault detection to ensure the reliability and cost-efficiency of electronic devices. Fault detection within this sector entails collaboration among multiple stakeholders, including Original Equipment Manufacturers (OEMs), Integrated Device Manufacturers (IDMs), wafer foundries, and software providers. A common challenge is the reluctance to share sensitive design data centrally, which is essential for building traditional machine learning models. To overcome these challenges, this paper introduces an innovative fault detection model that leverages Federated Learning (FL) and Explainable AI (XAI). FL's decentralized approach enhances model learning across multiple nodes without requiring the pooling of sensitive data, thus preserving data privacy. Concurrently, XAI ensures that the developed models maintain transparency and trustworthiness, even when trained on distributed datasets. This FL-based fault detection model permits stakeholders to train ML models on node-specific data without centralizing sensitive information. It accommodates heterogeneous and asynchronously-stored data, diverse machine learning models, and nodes with varying capacities and data volumes. By addressing the opacity of deep learning models, FL and XAI unveil their predictive behaviors in identifying semiconductor faults. Empirical results, obtained using a public dataset, demonstrate a significant improvement in defect identification precision, achieving an exceptional test accuracy of 98.78%. These findings underscore the potential of the proposed approach to transform fault detection in semiconductor manufacturing, thereby enhancing the reliability and efficiency of the production process.

INDEX TERMS Classification algorithms, data privacy, deep learning, explainable artificial intelligence, trusted AI, semiconductor materials.

ACRONYM TABLE

Acronym	Definition
AI	Artificial Intelligence.
FL	Federated Learning.
XAI	Explainable Artificial Intelligence.
WBM	Wafer Bin Maps.
CNN	Convolutional Neural Networks.
LIME	Locally Interpretable Model Agnostic Explanations.
GradCAM	Gradient Weighted Class Activation Mapping.

The associate editor coordinating the review of this manuscript and approving it for publication was Ye Zhou.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

EDA	Exploratory Data Analysis.
PCA	Principal Component Analysis.
t-SNE	t-distributed Stochastic Neighbor Embedding.
FedAvg	Federated Averaging.
DP	Differential Privacy.
KABLS	Knowledge Augmented and Broad Learning System.
SVID	Status Variable Identifications.
ART	Adaptive Resonance Theory.
DBMNN	Deep Multibranches Neural Network.

I. INTRODUCTION

Semiconductor manufacturing is essential to the progress of modern technology, with flaw detection during the design stage being vital for ensuring quality and safety. As given below, the challenges in detecting defects within semiconductors are identified and methods for addressing these challenges are discussed. It then proceeds to define research objectives, specifically concentrating on the use of Federated Learning (FL) and Explainable Artificial Intelligence (XAI) techniques to improve the defect detection processes in semiconductor design.

A. SEMICONDUCTOR MANUFACTURING AND ITS IMPORTANCE

Semiconductor manufacturing plays a vital role in shaping the landscape of modern technology, serving as the backbone for an excess of electronic devices that spread through every aspect of our daily lives [1]. The complicated and highly controlled processes involved in the process of semiconductor fabrication, such as lithography, etching, doping, and metallization, contribute to the creation of advanced electronic devices [2]. These devices, generally made from materials like silicon, are the foundation of microprocessors, memory chips, and integrated circuits that power an array of devices, ranging from smartphones and computers to highly advanced medical instruments and automotive systems. The constant effort to make semiconductors tiny and more powerful has propelled technology forward, encouraging new ideas and driving rapid developments in the electronics industry [3]. As we stand at the forefront of the next wave of technological development, understanding the details and significance of semiconductor manufacturing is vital for researchers, engineers, and policymakers alike, as they cross through the path toward a more connected and technologically advanced future.

B. IMPORTANCE OF IDENTIFYING FAULTS AT THE DESIGN LEVEL

Identifying faults at the design level in semiconductor manufacturing is vital for confirming the reliability, efficiency, and cost-effectiveness of the final product [4]. Detecting and addressing potential defects early in the design phase helps prevent expensive errors from propagating throughout the manufacturing process [5]. This prevention ultimately saves time and resources. By conducting detailed design verification and validation, semiconductor professionals can identify and rectify issues related to functionality, performance, and power consumption, confirming that the final product meets the intended specifications. Early fault detection also improves the overall quality of semiconductor devices, reducing the probability of defects that could lead to product failures, recalls, or compromised functionality [6].

C. THE CHALLENGES OF DETECTING FAULTY SEMICONDUCTORS AND THE SIGNIFICANCE OF ADDRESSING THIS ISSUE.

Faulty design in semiconductor manufacturing can lead to significant challenges and performance issues in electronic devices. One common issue arises from design flaws that result in insufficient power distribution, leading to irregular power delivery across the integrated circuits [7]. This can cause localized heating and, in some exceptional cases, thermal runaway, compromising the reliability and lifetime of the semiconductor components. Another serious concern is the existence of design errors that may result in signal integrity issues, such as crosstalk and electromagnetic interference, affecting the overall functionality of the semiconductor device [8]. In addition, if the arrangement and spacing of components are not carefully considered, it can lead to increased susceptibility to manufacturing defects, including short circuits and open circuits [9]. These faults not only threaten the efficiency of the semiconductor but can also contribute to higher production costs and increased rates of product failures after being sold. Fault detection in semiconductor manufacturing faces significant challenges, particularly due to the sensitive nature of the data, which restricts sharing across different organizations and facilities. Ensuring data privacy while enabling effective collaborative model training is critical to overcoming these challenges and improving defect detection accuracy.

D. OBJECTIVES AND CONTRIBUTIONS OF THE RESEARCH.

Detection of semiconductor faults at the design stage is crucial in ensuring their reliability and functionality in electronic devices. Traditional approaches encounter challenges around centralized model training and explainability in decision-making. We propose a novel approach that combines FL and XAI techniques to address these challenges and enhance the processes behind latent fault detection in semiconductor designs. The key contribution of the study relies on the innovative fusion of FL and XAI, paving the way for a new era of intelligent and transparent fault detection processes in semiconductor design

1) OVERCOMING LIMITATIONS OF CENTRALIZED MODEL TRAINING

FL facilitates collaborative learning across distributed devices, enabling model training without compromising the confidentiality of sensitive design data. This approach fosters

a secure and efficient environment for leveraging diverse datasets, which in turn enhances fault detection capabilities.

2) UTILIZING COLLECTIVE INTELLIGENCE OF DIVERSE DATASETS

By harnessing FL, the research capitalizes on the collective intelligence derived from diverse datasets. This collective intelligence enhances the fault detection capabilities beyond what individual datasets could achieve alone, thereby improving the overall robustness of semiconductor designs.

3) INTEGRATING EXPLAINABLE AI FOR TRANSPARENCY AND INTERPRETABILITY

XAI is integrated to address the critical need for transparency and interpretability in AI-driven fault detection systems. This integration ensures that the decisions made by the fault detection models are not only accurate but also explainable, providing designers with insights into the underlying factors contributing to detected faults.

E. STRUCTURE OF THE STUDY

The paper is structured into five main sections. Section I serves as the introduction, establishing the foundational framework for the research while also delineating the gaps present in the current body of knowledge. In Section II, a comprehensive review of related works is presented, emphasizing the identified research gaps for further exploration. Following this, Section III delves into the proposed framework of synergy of FL and XAI for fault detection in semiconductor manufacturing. Here, the methodology and intricacies of the proposed approach are discussed in detail. In Section IV, the focus shifts to the presentation and analysis of results obtained through the application of the proposed framework, accompanied by an exploration of potential future directions for research and development in this field.

II. RELATED WORK

The research article recommends a data-driven method for fault detection and diagnosis in semiconductor manufacturing [10]. The authors focus on detecting key status variable identifications (SVIDs) and key processing time and steps for fault detection. They have used the random forests method to analyze the importance of SVIDs and k-means clustering to identify the key SVIDs. They also used ensemble models built on k-nearest neighbours and naive Bayes classifiers for classifying wafers as normal or abnormal. The authors conduct an experimental study using thin film data in semiconductor manufacturing to test their framework. The results show that their proposed framework effectively detects abnormalities and provides valuable insights about SVIDs and corresponding processing time and steps. The study recommends that data-driven methods using machine learning methods are suitable for fault detection and diagnosis in semiconductor manufacturing.

Another interesting work proposed by authors in [11] gives a data-driven approach for fault detection and diagnosis in semiconductor manufacturing using image processing techniques and the Fourier transform. The authors find key parameters that have a significant influence on wafer quality by analyzing raw trace data and computing Fisher's criterion ratios. The authors transformed the raw trace data into 2D images and applied texture analysis with the Fourier transform to detect defective wafers. The results demonstrate that their approach successfully identifies key parameters and detects wafer defects. The proposed approach can be used for advanced process control and improving production yield in semiconductor manufacturing.

The research work in [12] discusses the challenge of identifying process related failures in semiconductor manufacturing due to the increasing complexity of wafer bin maps (WBM) patterns. The authors propose a knowledge-based intelligent system for WBM defect diagnosis and yield improvement in wafer fabrication. The system comprises a graphical user interface, a WBM clustering solution, and a knowledge database. The WBM clustering method integrates spatial statistics test, cellular neural network (CNN), adaptive resonance theory (ART) neural network, and moment invariant (MI) to group different patterns effectively. An interactive conversation-based interface is developed to present the actual root causes in the order of similarity matching and record the diagnosis know-how from domain experts into the knowledge database. The proposed solution has been implemented and tested in a leading semiconductor manufacturing company in Taiwan.

The research work proposed by authors in [13] discusses about defect detection in semiconductor manufacturing systems, particularly in identifying mixed-type defects in integrated circuit wafers. The authors introduce a knowledge augmented broad learning system (KABLS) with a knowledge module and broad selective sampling module to provide a multichannel selective sampling network to decouple the mixed-type defects. The model uses pre-trained deformable convolution units in each channel to extract the feature of a fixed single-type defect. The knowledge module is designed to activate the candidate network channel by pre-detection of wafer maps, and the broad selective sampling module separates a mixed-type defect into several basic defects for accurate identification. The authors evaluated the proposed KABLS against five other state-of-the-art models and found that KABLS outperformed the other models in terms of accuracy. Numerical experiments were conducted on a mixed-type wafer map dataset, and the results showed that KABLS has the maximum classification accuracy when the learning rate is 0.0001 and the number of training epochs is 100.

Another significant research work [14] is about defect detection in wafer semiconductor surface inspection using deep convolutional neural networks (CNNs). The authors introduce a novel method that combines a fully convolutional network with region proposal network and deep

multibranches neural network to detect and segment wafer defects. The proposed method uses a two-stage framework: the first stage generates region proposals using a region proposal network (RPN) to locate potential object areas, and the second stage performs the segmentation using a deep multibranches neural network (DMBNN). The proposed method also uses a feature pyramid network with atrous convolution (FPNAC) to generate feature maps at different scales, with which the RPN can provide both classifiers and bounding boxes for each region proposal. The proposed method was compared to other state-of-the-art methods, and the results show that it outperformed them in terms of mean pixel accuracy and mean intersection over union metrics.

The authors in [15] discuss about a defect classification method for semiconductor manufacturing on extremely small datasets using geometrically varied synthetic data and pretrained deep neural networks (DNNs). The authors introduce a solution that consists of three components: an image capturing unit, a data processing unit, and a visual display unit. The data processing unit uses a two-stage process: model generation and defect identification. To generate geometrically varied synthetic data, the proposed solution retrieves foreground images from the current dataset and applies geometric transformations on each image, such as rotation, scaling, and translation. The authors also propose a pretrained model selection method that evaluates pre-trained DNN models based on their structural value and hashing differences. The model with the highest total average evaluation score is selected as the best suitable pretrained model for the defect classification task. The proposed solution was evaluated on an extremely small dataset, and the results show that the proposed method outperformed traditional machine learning methods in terms of classification accuracy. The authors conclude that the proposed solution provides a practical and effective way to perform defect classification on extremely small datasets in semiconductor manufacturing.

Another research [16] aims to develop an intelligent system that can recognize defect spatial patterns on semiconductor wafers using a neural network approach. The ART1 neural network architecture was adopted for this purpose, and actual data obtained from a semiconductor manufacturing company in Taiwan were used in experiments with the proposed system. Comparison between ART1 and another unsupervised neural network, self-organizing map (SOM), was also conducted. The results show that ART1 architecture can recognize the similar defect spatial patterns more easily and correctly. The system was designed to detect a greater number of different spatial patterns on a wafer. The research concludes that the ART1 neural network is highly desirable for detecting and recognizing spatial defect patterns in semiconductor fabrication.

A. RESEARCH GAPS

Existing methods [12], [13], [14], [15], [16] of fault detection in semiconductor systems are often found difficult due to the inherent complexity of these systems and the limitations of traditional approaches. Many existing methods rely on centralized data processing, which can be inefficient and may raise worries about data privacy and security from the industry perspective. Furthermore, these approaches may not adequately capture the details of faults that can arise during the design phase, leading to incomplete or inaccurate detection. FL can be a convincing alternative by enabling collaborative model training across multiple decentralized clients while keeping raw data localized, thus addressing privacy concerns. In the context of semiconductor fault detection, FL makes use of the distributed knowledge obtained from various clients, including manufacturers, designers, and testers. This distributed way allows FL models to learn from a diverse range of data sources, including different manufacturing processes, design variations, and environmental conditions. By aggregating insights from these sources, FL can offer more robust fault detection capabilities, improving the reliability and effectiveness of fault detection during the design phase of semiconductor systems. Furthermore, integrating XAI techniques further improves the effectiveness of fault identification by providing transparent and interpretable insights into the detection process, thereby ensuring greater reliability and trustworthiness in semiconductor fault detection methodologies.

III. PROPOSED METHODOLOGIES

This section outlines the anticipated procedure for our research, beginning with a summary of the work and emphasising the significance of XAI and FL. Explaining the Deep Learning models for categorization and explore the importance of FL, including ideas such as Federated Averaging and Differential Privacy. Furthermore, we analyse the significance of XAI and present a comprehensive outline of techniques including LIME and GradCAM. Moreover, we provide a detailed description of the dataset utilised in our work and present a clear explanation of the recommended methodology for conducting experiments.

A. OVERVIEW OF THE WORK

The problems of defect classification are a persistent problem for the semiconductor sector, which is the backbone of contemporary technology. Defect identification and categorization are extremely difficult tasks due to the complex and integrated structure of semiconductor manufacture processes. Numerous flaws can affect semiconductor wafers, such as lithographic mistakes, etch anomalies, and material contaminants. The performance and dependability of the finished product can be greatly impacted by these flaws, so it is critical to accurately detect and classify them in order to uphold strict quality and yield standards.

The sheer volume and variety of faults in semiconductors presents one of the main categorization challenges. According to a study [17], conventional detection methods are insufficient for semiconductor wafer defects since they might vary widely in size, shape, and kind. Additionally, as a result

TABLE 1. Related works.

Ref. No	Technique Used	Results	Challenges
[10]	Random forests, k-means clustering, ensemble models (k-nearest neighbors, naive Bayes classifiers)	Effective fault detection, identification of key SVIDs, classification of normal/abnormal wafers	N/A
[11]	Image processing, Fourier transform, texture analysis	Successful identification of key parameters, detection of wafer defects	N/A
[12]	Knowledge-based system, WBM clustering (spatial statistics test, cellular neural network, adaptive resonance theory, moment invariant)	Effective WBM defect diagnosis, yield improvement, interactive user interface	Handling increasing complexity of WBM patterns
[13]	Knowledge augmented broad learning system (KABLS), pretrained deformable convolution units, broad selective sampling module	Outperformed other models in accuracy, decoupling of mixed-type defects, maximum classification accuracy	Selection and integration of suitable modules, optimization of learning parameters
[14]	Deep convolutional neural networks (CNNs), fully convolutional network, region proposal network (RPN), deep multibranches neural network	Outperformed state-of-the-art methods in defect detection, improved mean pixel accuracy	Optimizing network architecture and training parameters, computational complexity
[15]	Pretrained deep neural networks (DNNs), geometrically varied synthetic data	Outperformed traditional machine learning methods, practical and effective defect classification	Developing suitable techniques for synthetic data generation, selecting appropriate pretrained models
[16]	ART1 neural network	Improved recognition of defect spatial patterns, greater number of detected spatial patterns	Comparison with other neural network approaches, generalization to different datasets

of the semiconductor industry's rapid growth, feature sizes are getting smaller and smaller. This makes defect detection more difficult because previously trivial faults are now crucial failure points.

The enormous volume of data produced throughout the semiconductor manufacturing process is another major obstacle. Modern semiconductor fabrication plants are outfitted with a multitude of sensors that produce vast amounts of data [7]. This data must be efficiently evaluated in order to discover defects. Conventional defect classification techniques are labor-intensive and prone to mistakes and inconsistencies since they frequently rely on manual inspection or basic computational approaches.

Defect classification is further complicated by the dynamic nature of semiconductor processes, which involve constant adjustments to production conditions and process recipes in order to maximize performance. According to a study [18], this dynamic environment necessitates the use of adaptive classification methods that may change as process conditions do.

FL represents a transformative approach in the realm of semiconductor defect classification, addressing some of the most significant challenges faced by the industry. By enabling data to remain on local devices while aggregating model updates centrally, FL offers a solution to the privacy and security concerns associated with transferring large volumes of sensitive manufacturing data over networks. This decentralized learning paradigm allows for the collection and analysis of a diverse and comprehensive dataset from various points in the semiconductor manufacturing process without compromising on data confidentiality. Furthermore, FL facilitates the creation of more robust and generalized models by leveraging data from a multitude of sources, each with potentially unique defect characteristics and manufacturing environments. This approach not only enhances the accuracy of defect classification models but also adapts to the dynamic nature of semiconductor processes by continuously learning from new data. Incorporating FL into semiconductor defect analysis empowers the industry to harness the full potential of AI while addressing data privacy, security, and model generalization challenges head-on, paving the way for more resilient and efficient manufacturing processes.

These difficulties highlight the necessity for sophisticated techniques in semiconductor defect classification that can manage enormous datasets, accommodate a wide range of dynamic defect forms, and yield accurate and consistent classification results. Herein lies the potential benefit of combining XAI techniques with deep learning models. Deep learning models are useful for identifying and categorizing a variety of semiconductor flaws because of their capacity to analyze massive amounts of data and discover intricate patterns. However, the often 'black box' nature of these models raises concerns about interpretability and trustworthiness, which is where XAI methods come into play. By providing insights into the decision-making process of deep learning models, XAI methods enhance transparency and reliability, making them indispensable tools in the quest for efficient and accurate semiconductor defect classification.

B. IMPORTANCE OF EXPLAINABLE ARTIFICIAL INTELLIGENCE AND FEDERATED LEARNING IN ENHANCING TRANSPARENCY AND COLLABORATIVE LEARNING IN DEEP LEARNING MODELS

Deep learning models can handle enormous datasets and complex manufacturing data patterns, making them ideal semiconductor defect classification tools. However, the

training process is not truly centralised, making it look like a solo process at edge systems. FL links traditional centralised training to decentralized learning sharing via various approaches, and XAI bridges the gap between powerful computer capabilities and human interpretability, which is crucial in the centralized training phase due to model inexplicability.

- *FL for Collaborative Learning* offers a novel approach to data utilization and model training. Data privacy and proprietary issues are crucial in semiconductor production. FL lets industrial units train models without sharing raw data, boosting privacy and using several datasets for better learning.
- *Encouraging Human Oversight* Human expertise is still important in semiconductor quality assurance. XAI connects AI models with experts. XAI allows specialists to assess outcomes by providing insights into the model's logic, assuring consistency with real-world knowledge and expectations [19].
- *Improving Debugging and Model Improvement*To improve deep learning models and build trust, one must understand how they work. A study [20] suggests that XAI can identify areas for model development, enabling researchers to optimize accuracy and performance.
- *Compliance and Ethical Considerations* Standards and ethics are crucial in highly regulated industries like semiconductor fabrication. XAI helps deep learning models meet legal standards by making their operations clear and logical. A study in [21] addressed this issue in the perspective of accountable AI systems.
- *Transparency in Decision Making* Transparency in deep learning model decision-making is a major aspect of XAI. An extended study [22] emphasizes the need of explainability in AI for understanding outcomes and ensuring reliability in high-stakes sectors like semiconductor production. XAI helps engineers and decision-makers understand why a model flags a wafer area as bad, boosting automated system trust.
- *Combining XAI and FL* Combining XAI and FL addresses two key issues: improving model performance without compromising data privacy, and understanding AI judgments using XAI. Using XAI to FL-trained models lets stakeholders see how data from diverse sources affects model predictions and trust them.
- *Enhancing Industry 4.0* In the context of Industry 4.0, the synergy of XAI and FL fosters a more collaborative, transparent, and efficient approach to AI-driven problem-solving in semiconductor defect classification, paving the way for smarter and more reliable manufacturing processes.

C. INTRODUCTION TO DEEP LEARNING MODELS FOR SEMICONDUCTOR DEFECT CLASSIFICATION

Deep learning, a subset of machine learning, has transformed image analysis and complicated pattern identification. Artificial neural networks, inspired by the brain, underpin deep learning. Machines can process, classify, and cluster raw data using these networks to identify patterns and interpret sensory input. Neural networks are powerful because they process data in layers. This multilayered method lets the system learn "deep" from data across abstraction levels [23].

A critical aspect of deep learning models is their ability to learn from unstructured data without explicit supervision. This capability largely stems from the concept of a feature hierarchy. In this hierarchy, intermediate layers of the network extract higher-level features (such as objects and shapes in images) from basic lower-level features (like edges and textures). This method of hierarchical learning is particularly beneficial in fields like semiconductor defect detection, where identifying the slightest irregularities in patterns can be crucial. The hierarchical learning approach enables these models to excel in tasks requiring fine-grained recognition and classification, demonstrating deep learning's profound impact on technology and industry [24].

Defect classification is a crucial quality control measure in the semiconductor manufacturing industry, requiring advanced analytical models with high-precision identification capabilities. We pursue this goal by combining deep learning architectures supported by strong mathematical frameworks, each of which makes a distinct contribution to the identification and categorization of semiconductor faults.

The Residual Network (ResNet) model, and specifically the ResNet152 variant, addresses the challenge of training very deep neural networks. In their groundbreaking article [25], the authors introduced ResNet, which utilizes residual learning to ease the training of networks that are substantially deeper than those used previously. ResNet152, with its depth of 152 layers, uses skip connections or shortcuts to jump over some layers. These connections help in combating the vanishing gradient problem, allowing for the training of very deep networks. ResNet152's ability to learn from a considerably increased depth of layers contributes to its high accuracy, as evidenced in its performance in the ImageNet dataset.

The concept of ResNet is pivotal, the model is designed to learn residual functions with reference to the layer inputs, as expressed by the formulation as given by the eq. (1)

$$F(x) + x \tag{1}$$

where $F(x)$ is the residual mapping to be learned and the employment of such skip connections allows the training of much deeper networks and addresses the vanishing gradient problem. In practice, this facilitates the detection of a broad spectrum of defects, ranging from minuscule particulate contamination to pattern irregularities that are discernible only through deep layered analysis.

ResNet was particularly chosen for semiconductor wafer defect classification task as it has proven results in terms of multivariate time-series data for fault detection in semiconductor manufacturing equipment as proven by the authors in [26] who have particularly used 1D ResNet for Multivariate

Fault Detection in Semiconductor Manufacturing Equipment and have gained a high F-score of 0.9708, Thus, making ResNet a good choice for Image Classification for Semiconductor Wafer Defects.

DenseNet, as proposed by the authors in [27] which is characterised by its unique connectivity pattern is used by the authors in [28] and explores an innovative approach of using Deep Learning for recognition of Inline Defects for the production of semiconductors wherein they have explored evaluation for 9 different defect types and 14 production steps, for which they have used DenseNet201 which had given 96.04% for multiclass classfication, thus giving a motivation to use DenseNet for the current task at hand of Defect Classficication for Semiconductor Wafer Maps.

MobileNet, as introduced in [29] which is a class of efficient models designed for mobile and edge device applications was yet another choice for the task of classifying semiconductor faults as the authors in [30] explore various models for classfication of 9 defects in semiconfuctor wafer maps, where MobileNet provided an accuracy of 97.9%, hence making it as one of the choices for the said task.

This flexibility provided by the InceptionV3 architecture which was proposed by the authors in [31] which features an architecture with symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenation layers, dropouts, and fully connected layers is essential in semiconductor defect classification, where defects may not conform to a single scale or pattern. By including these mathematical concepts into the models, their capacity to identify the intricate and minute flaws typical of semiconductor wafers is improved. The models' unmatched precision in fault categorization is made possible by their exact, mathematically based examination, which also ensures the dependability of semiconductor components in real-world applications.

D. FEDERATED LEARNING: OVERVIEW AND SIGNIFICANCE

FL, a new machine learning method, embodies Industry 4.0's cooperation. While maintaining local data, it generates sophisticated prediction models using data from several devices or production sites. Competitive advantage in semiconductor production depends on trade secrets and sensitive data. [32]. Fast, intricate semiconductor production defines the industry. Even little defects can cause large financial losses and reduced electrical device reliability; therefore, defect categorization must be precise. FL enters this business to conserve node data and increase forecast accuracy by learning from a large, scattered data network [33].

A central server coordinates learning among several nodes with local datasets in FL architecture. Central servers initialise and deliver global models to nodes. Each node then trains the model on its local data. The server then receives parameter changes, or gradients. Privacy and security are addressed by keeping raw data on the node [34] A server aggregates node updates to change the global model. FedAvg is usually used for aggregation. The server calculates a weighted average of updates based on local dataset size. Data-rich nodes will correspondingly change the model. Differential privacy meets rising privacy needs in this architecture, it hides node contributions and avoids reverse-engineering model changes to reveal private data by carefully introducing stochastic noise [35].

Early silicon wafer flaw diagnosis improves semiconductor yield and quality. FL lets firms train fault categorization models without moving data between facilities and geographies. The decentralised method's privacy, security, and defect pattern variation leads to more accurate and generalised models [36]. Industry 4.0 transcends tech. Need sustainable, privacy-preserving, collaborative growth solutions. FL supports this movement by ensuring that the business flourishes with a paradigm that safeguards personal data and uses communal knowledge for society.

Algorithmically, FL as a whole can be formulated as in Algorithm 1

Algorithm 1 Federated Learning

Initialization: A Global model with parameter θ is initiated.

Local Training: The Global model with the set parameter is sent to a subset of clients with each device Ki updating the model based on the local data for the client, hence resulting in a local model update as in eq. (2)

$$\delta\theta_k = LocalTraining(\theta, Data_k) \tag{2}$$

Global Aggregation: The Local Client Model Updates are sent back to the central server or the global server and are aggregated to update the global model. Specifically, Weighted Averaging is used here as given in eq. (3)

$$\theta_{\text{global}} = \theta + \sum_{i=1}^{i=K} w_i \cdot \Delta\theta_i \tag{3}$$

where w_i is the weight that is set based on the number of samples or the size of the dataset on device Ki.

Iteration: Steps Local Training & Aggregation are iteratively performed till convergence or a set number of rounds

E. FEDERATED LEARNING WITH DIFFERENTIAL PRIVACY

The FL paradigm is based on the Federated Averaging (FedAvg) algorithm and Differential Privacy (DP). This combination strengthens the privacy context that these activities take place in, while also improving the collaborative learning process across different nodes. [33]

1) FEDERATED AVERAGING

The goal of FedAvg is to create a logical global model by combining updates from a dispersed network of clients,

each with a local dataset. This method works effectively in situations when data cannot be centralized because of legislative restrictions, privacy concerns, or bandwidth limitations. These are problems that the semiconductor industry deals with on a regular basis.

The Federated Averaging Process can be described in the steps as given in Algorithm 2.

Algorithm 2 FL With Differential Privacy

Input:

- K, the total number of nodes participating in the training
- B, the local mini-batch size
- E, the number of local epochs
- η, the learning rate
- σ, the standard deviation of the Gaussian noise for differential privacy
- S, the privacy budget per iteration

Procedure:

Initialize the global model weights w_0

For each round $t = 1, 2, \ldots$ do

- Server selects a random subset of nodes S_t
- For each node $k \in S_t$ in parallel do
  - $w_{t+1}^k \leftarrow \text{LocalTraining}(k, w_t, B, E, \eta)$
- $\Delta w_t \leftarrow \text{Aggregate}(\{w_{t+1}^k\}, k \in S_t)$
- $\Delta w_t \leftarrow \Delta w_t + \text{GaussianNoise}(0, \sigma^2)$
- $w_{t+1} \leftarrow w_t + \eta \cdot \Delta w_t$
- If PrivacyBudget(S, t) exhausted then
  - break

Output:

- The final global model weights w_t

The aggregation step at the server is defined as in eq. (4)

$$w_{t+1} = w_t + \frac{\eta}{N}\left(\sum_{k=1}^{K} n_k \Delta w_t^k\right) \quad (4)$$

where w_t is the global model weights at the t iteration, η is the learning rate, N is the total number of data points distributed across node k, and Δw_t^k is the update from node k.

2) DIFFERENTIAL PRIVACY

Through the introduction of a degree of randomness into the data or model changes that are transmitted during the FL process, Differential Privacy provides an extra layer of privacy protection. This provides strong privacy assurances by ensuring that any single data point does not significantly affect the conclusion of the aggregate computation.

By including precisely calibrated noise in the updates, DP can be applied to the FedAvg model during the aggregation phase. Typically, this entails boosting the weighted average of the model updates with Gaussian or Laplacian noise which can be calculated as in eq. (5)

$$w_{t+1} = w_t + \frac{\eta}{N}\left(\sum_{k=1}^{K} n_k \Delta w_t^k + \mathcal{N}\left(0, \sigma^2 I\right)\right) \quad (5)$$

where $\mathcal{N}\left(0, \sigma^2 I\right)$ is the Gaussian Noise with mean 0 and variance proportioning to sensitivity of the aggregation algorithm to the individual updates. The variance has been chosen based on the desired privacy budget, quantifying the allowable privacy loss.

F. EXPLAINABLE ARTIFICIAL INTELLIGENCE AND ITS SIGNIFICANCE

XAI addresses the interpretability issue in deep learning models. Despite their effectiveness in many sectors, deep learning models are frequently called "black boxes," having no visibility into how they make judgements. In sensitive fields like healthcare, finance, and autonomous systems, trust, diagnosticity, and compliance are crucial. XAI aims to make model outputs understandable to human specialists [21].

Not enough can be said about XAI in deep learning. XAI ensures that AI models are unbiased and fair, simplifies model debugging and enhancement, and helps comply with regulations by explaining model judgements. XAI is both technological and legal because the EU's General Data Protection Regulation (GDPR) requires automated systems to explain their judgements [37].

XAI approaches include various explainability aspects. Gradient-based approaches, such Gradient-weighted Class Activation Mapping (Grad-CAM), highlight prediction-critical input image regions and provide heatmaps to illustrate CNN results. Others, like Local Interpretable Model-agnostic Explanations, seek model-agnostic explanations. LIME perturbs input data and observes output to approximate the model locally [38].

Approaches like SHAP (Shapley Additive Explanations) although being an excellent tool for generating explainable insight, it fails to provide high visual interpretability in terms of overlapping the explanations received from the same. Whereas, Grad-CAM and LIME align more to the fundamental aspect of the study, which being the need to visually interpret which parts of the images contribute to model decisions.

The use of XAI techniques in the field of semiconductor defect categorization may enhance the dependability and credibility of automated inspection systems. Engineers can improve the production process and quality control procedures to lower the frequency of faults by knowing the reasoning behind specific classifications.

G. OVERVIEW OF APPLIED EXPLAINABLE AI METHODS

Several XAI techniques have been developed as a result of the machine learning community's pursuit of openness. These are essential resources for deciphering intricate deep learning models, such as those used in the categorization of semiconductor defects. XAI is a set of processes and

methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. The methods that are applied for Semiconductor Defect Classification are described below.

1) LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

LIME is a XAI technique that produces interpretable and local surrogate models that roughly match the original model's predictions in an effort to demystify any black-box classifier's decisions. The idea behind LIME is that although a complicated model's decision boundary may be multidimensional and nonlinear, it may be linear and simple close to a specific instance that needs to be explained [39].

Every complicated model is assumed to be linear at the local level in the theoretical foundation of LIME. To put it another way, for each prediction, there is a simple model that may accurately represent the behavior of the complicated model and be used to understand the reasons for a particular prediction. LIME's interpretability stems from its simplicity; it employs human-readable models such as decision trees, rule lists, and linear regression.

Mathematically, LIME explain the instance x by learning an interpretable model g belonging to a class G (e.g., linear models or decision trees), where g is assumed to be locally faithful to a classifier f, hence meaning that g approximates f precisely when x is close to the instance that is getting explained. The faithfulness of g to f in the vicinity of x is measured by a locality-based kernel π_x which is defined as the width of the neighborhood. The explainable model g is obtained by the following optimization model as in eq. (6)

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{6}$$

where, $\xi(x)$ is the explanation for instance x, $\mathcal{L}$ is a loss function that measures the unfaithfulness of g in approximation of f in the locality defined by π_x and $\Omega(g)$ is the measure of the complexity of the model g. The primary goal is to minimize $\mathcal{L}$ while also keeping g as simple as possible which is controlled by the complexity factor.

This optimization problem is solved by generation of a new dataset consisting of perturbed samples around x and the corresponding predictions by f which are weighted by π_x, which typically is an exponential kernel defined on some distance metric and the model g has been trained on this weighted dataset as given in eq. (7)

$$\pi_x(z) = \exp\left(-\frac{d(x, z)^2}{2\sigma^2}\right) \tag{7}$$

Here, z is the perturbed sample, $d(x, z)$ is the distance between the original instance x and z and σ being the kernel width control parameter. It can algorithmically be represented as shown in Algorithm 3.

LIME's ability to provide local explanations makes it a powerful tool, especially when the overall model behavior is too complex to be comprehensively explained, or when explanations are needed for individual predictions rather than the model as a whole.

Algorithm 3 Procedure for LIME Method

Select an Instance for Explanation: Choose the instance x for which you want to explain the prediction.
Perturbation of Data: Generate a new dataset by perturbing the features of x, creating many synthetic samples (z) around x. This dataset would be covering the locality around x where the explanation is aimed to be accurate.
Prediction using original model: Use the complex model f to predict outcomes for all the perturbed samples. This step captures how the prediction changes with slight variations in the input around the instance x.
Weighting the Synthetic Samples: Assign weights to the synthetic samples based on their proximity to instance x. This is done by the kernel π_x as defined in eq. (7).
Learning the interpretable model: The trained model on the dataset of perturbed samples and their corresponding weights is fit into an optimization equation as shown in eq. (6) where the goal is to find the model g to reduce the Loss function $\mathcal{L}$ indicating how well g approximates f in the local region defined by weights, while also considering the complexity of $g(\Omega(g))$. The optimization is formulated as in eq. (6)
Explanation Generation: The interpretable model g now serves as the explanation for the prediction of x by the complex model f The parameters of g (e.g., coefficients in a linear model) indicate the importance of each feature for the prediction of x.
Interpretation and Analysis: Analyze the interpretable model g to understand which features contributed most to the prediction and how they influenced it. This step is crucial for gaining insights into the decision-making process of the complex model f in the local vicinity of x

2) GRADIENT ACTIVATED CLASS ACTIVATION MAPPING (GRAD-CAM)

Grad-CAM is a technique that visualizes the portions of the input that are crucial for predictions on particular classes, thus increasing the transparency of CNN-based models. It works particularly well for tasks like localization and image classification and can be implemented on any CNN-based architecture without the need for retraining or architectural modifications.

Grad-CAM relies on the hypothesis that high-level visual constructs are captured by CNN's higher-level convolutional layers. Information regarding each neuron's significance for the relevant choice is sent by the gradients that flow into these layers during the backpropagation phase. By utilizing these gradients, Grad-CAM produces a coarse heatmap that is equivalent in size to the convolutional feature maps. The

target class's key areas in the image are highlighted in this heatmap, also known as a class activation map.

The Mathematical Formulation of the Gradient Activated Class Activation Map can be described as an algorithm which is stated as in Algorithm 4.

Algorithm 4 Mathematical Explanation: Grad-CAM

Forward propagate the image through the network to obtain the class scores before the softmax layer.
Identify the **feature maps** A^k of the last convolutional layer.
Compute the **gradient** of the score for the class c, Y^c with respect to the feature maps A^k, $\frac{\partial Y^c}{\partial A^k}$.
Perform global average pooling over the width and height dimensions of the gradient to get the neuron importance weights α_k^c.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{8}$$

where Z is the number of pixels in feature map, and A_{ij}^k is the activation of feature map k at pixel location (i, j).
The Neuron Activation Weights are combined with **forward activation maps** to obtain the **class activation map**

$$L_{\text{Grad-CAM}}^C = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{9}$$

where ReLU is applied to only consider features that have a positive influence on the class of interest.

Grad-CAM thus generates a heatmap that can be overlaid on the input image to show the discriminative regions used by the CNN to identify that class. This is a powerful way to visualize and understand which parts of the image are deemed important by the neural network.

H. DATASET DESCRIPTION

The dataset used in this study is made up of high-resolution semiconductor wafer images that have been carefully selected to help enhance defect classification techniques. The dataset, which was first presented in [40], includes a wide variety of defect types that are essential to the production of semiconductors.

The collection consists of over 38,000 52 × 52 pixel resolution grayscale images distributed across 38 different classifications. Every class represents a distinct flaw or a set of faults, carefully identified to correspond with standard defect taxonomy in the semiconductor manufacturing industry. Standard preprocessing methods, such as grayscale normalization and image flattening, were used before the analysis to promote consistency throughout the dataset. This standardization process, which is carried out using Python's StandardScaler, guarantees that each pixel intensity will contribute equally to the results of any further analytical processes.

TABLE 2. Single type defect classes.

Name	Acr.	Description
Centre	C	Defective Die scattered in the centre of the Wafer Map
Donut	D	Defective Die from the centre of the Wafer Map in a Ring like configuration
Edge - LOC	EL	Localised clusters in the Wafer Map around the edges
Edge - Ring	ER	Ring Clusters around the edges in the Wafer Map
Near-Full	NF	Unusual Faults occuring all over the Wafer Map
Local	L	Localised Faults spread over the Wafer Map
Scratch	S	Narrow faults in an unusual manner in a long region
Random	R	Random Defect with no classification to any other defect and is without patterns

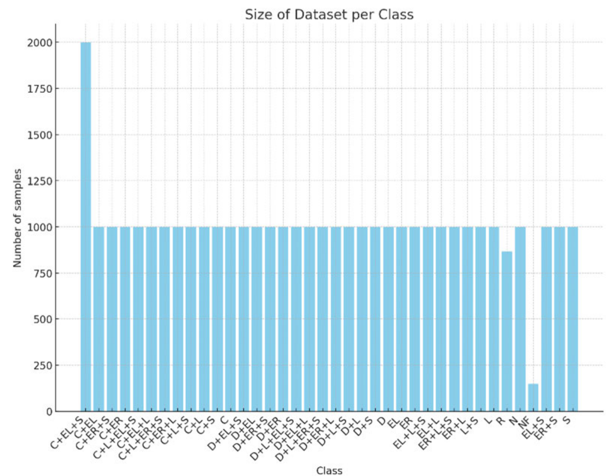

FIGURE 1. Distribution of dataset.

To find underlying patterns and distributions in the dataset, extensive EDA was carried out. One way to mitigate potential biases in model training was to disclose a balanced representation of fault kinds by statistical studies of class frequencies which is shown in the Fig. 1.

As observable in the Fig. 1, 38 different classes are present in the MixedWM38 Dataset indicating a multitude of errors and faults in the semiconductor wafers. The classes comprise of one fault-free pattern, eight single-defect patterns, thirteen two-mixed type patterns, twelve three-mixed type patterns, and four four-mixed type patterns. The dataset, as described earlier has images of 52 × 52 size, whose class information is one-hot encoded in the array in a numpy format. The table describes the base eight single-defect patterns, followed by the 38 different defect patterns.

In order to reduce the dataset's dimensionality while retaining as much of its variability as possible, Principal Component Analysis (PCA) was used to transform the dataset into a set of principal components. The results of the analysis showed that the top 50 principal components

TABLE 3. 38 Different defect patterns in MixedWM38 dataset.

No.	Single Defect	No.	Two-Mixed Defect	No.	Three-Mixed Defect	No.	Four-Mixed Defect
1	Normal	10	C+EL	23	C+EL+L	35	C+L+EL +S
2	Center (C)	11	C+ER	24	C+EL+S	36	C+L+ER +S
3	Donut (D)	12	C+L	25	C+ER+L	37	D+L+EL +S
4	Edge - LOC (EL)	13	C+S	26	C+ER+S	38	D+L+ER +S
5	Edge - Ring (ER)	14	D+EL	27	C+L+S		
6	LOC (L)	15	D+L	28	D+EL+L		
7	Near Full (NF)	16	ER+L	29	D+EL+S		
8	Scratch (S)	17	EL+S	30	D+L+S		
9	Random (R)	18	ER+S	31	D+ER+L		
		19	L+S	32	D+ER+S		
		20	D+ER	33	EL+L+S		
		21	D+S	34	ER+L+S		
		22	EL+L				

captured 30.22% of the variation overall, emphasizing the dataset's high dimensionality and underlying complexity.

The PCA-reduced data was then subjected to t-SNE, which produced a two-dimensional representation that made it easier to see the fundamental structure of the dataset. After 300 iterations, the t-SNE algorithm converged to a KL divergence of 3.846083 and produced a mean sigma of 3.696448. These measurements show how well the algorithm clusters related data points while preserving some degree of distinction between various defect categories.

The Fig. 2 that is attached shows the t-SNE visualization. It shows several clusters that correlate to the different types of defects in the dataset. Interestingly, regions in the picture indicate classes that have similar features, such 'ER' in 'C+ER+S' and 'C+ER+L,' highlighting possible relationships between specific problems. These findings are supported by the PCA projection in Fig. 3, which is also presented below. It shows clear clusters that imply some degree of association between classes that have similar defective components.

The sample of different defects are shown as in the Fig. 4:

I. PROPOSED SETUP

In this work, we present a FL setup with a distributed dataset of about 38,000 images across 38 classes, specifically designed for the purpose of semiconductor defect classification. We present the Explainable AI (XAI) techniques that are used to analyze the results, together with an overview of the FL framework and the model training procedure. Ten clients make up our federated system, and each one has a portion of the complete dataset. These clients are specific manufacturing facilities or inspection stations that include hardware for taking pictures that can capture images of semiconductors. In order to preserve data security and privacy, a FL strategy is suitable given the sensitive nature of the data and the proprietary processes it represents.

- Initialization: The global model is initialized by a central server. A Convolutional Neural Network (CNN) is a suitable option for image classification because of its ability to handle picture data. To maintain uniformity, the model architecture and hyperparameters are predefined and shared throughout customers.
- Local Training: Using the FedAvg algorithm, each client trains the model independently on its own dataset. The clients calculate changes to the model weights after completing multiple training epochs on their individual datasets.
- Communication: Each client (typically a manufacturing unit or inspection station equipped with sensors and imaging tools) sends updates to the central server after conducting local training on its model weights. These updates pertain solely to the parameters of the model, rather than the raw data gathered from the wafer inspection processes. To enhance the security of this sensitive information during transmission, the model weight updates can be encrypted. This ensures that proprietary or confidential information about the semiconductor manufacturing process is protected against potential interception or unauthorized access while maintaining the privacy of the data.
- Aggregation and Differential Privacy: These updates are combined into a new global model by the central server. Differential privacy is maintained during this aggregation process by introducing Gaussian noise into the updates, guaranteeing that the contributions from distinct clients cannot be identified.
- Global Model Update: The server uses the aggregated, differentially private changes to update the global model. It then returns the updated global model to the clients for the subsequent training cycle.
- Iteration: Until the model's performance converges or reaches a predetermined accuracy criterion, this process is repeated multiple times.
- Final Model Weights: The final model weights are exported as an H5 file, which is a common file format used to store neural network weights, after training is finished.

We use XAI approaches to interpret the federated model's decision-making process once it has been trained. Ensuring that the model's predictions are transparent and comprehensible to human specialists is the aim. The subsequent actions are performed:

- Model Interpretation: To provide mathematical and visual justifications for the model's predictions,

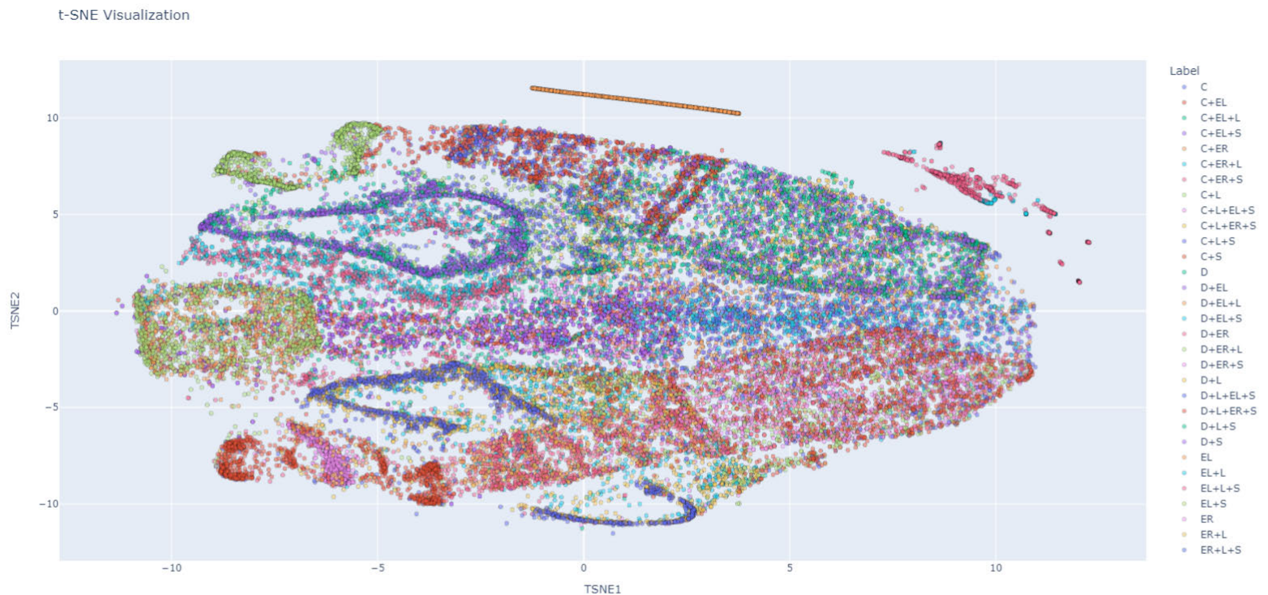

FIGURE 2. t-SNE visualisation.

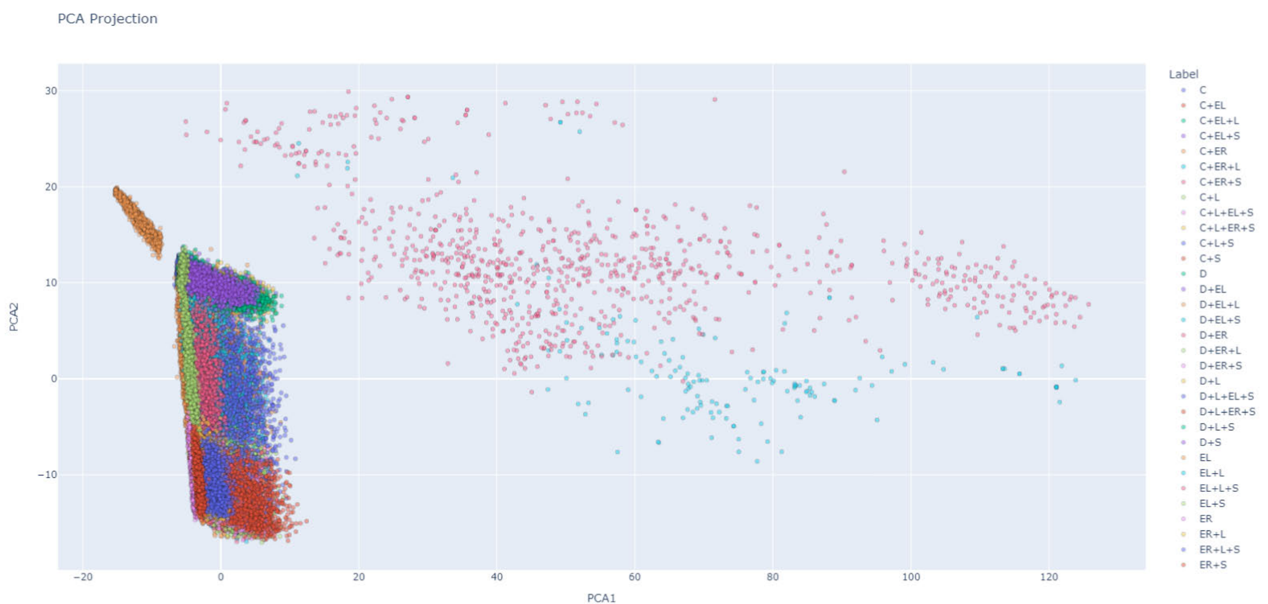

FIGURE 3. PCA projection.

we utilize a range of XAI techniques, including Grad-CAM, LIME, and Meaningful Perturbations. By emphasizing the areas of the pictures that have the most influence on the categorization choice, these techniques can shed light on the behavior of the model.

- Using XAI on Test Data: A series of test photos are subjected to the interpretability techniques, and the outcomes are examined in order to assess the model's effectiveness. We can check if the model is concentrating on the right patterns and features related to semiconductor faults using methods like heatmaps and feature significance scores.
- Assessment and Feedback Loop: The interpretations support a qualitative assessment of the model. A feedback loop is created where the knowledge collected can be used to improve the model or data collecting procedure if the XAI approaches show that the model is making judgments based on extraneous features or artifacts.
- Results: The results of the XAI techniques are recorded in thorough reports that give stakeholders intelligible explanations of the model's decision-making procedure. This can help in fine-tuning the model even further and informing choices about using it in real-world settings.

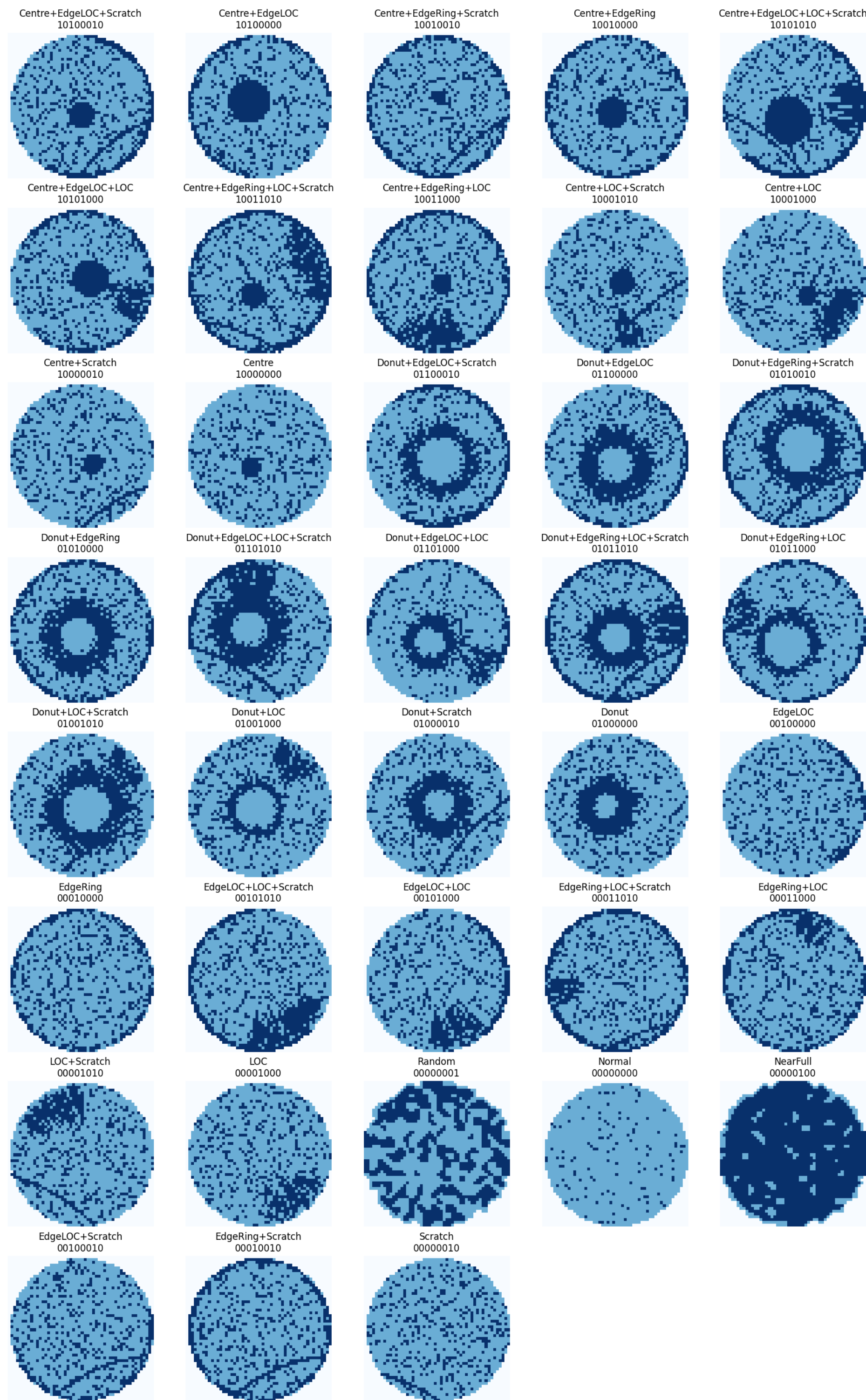

FIGURE 4. Various kinds of defects.

Designed to align with current industry standards, the fault detection model is capable of handling diverse semiconductor wafer data. This allows for the training and deployment of a robust global model. Industries can incorporate an MLOps Continuous Integration and Continuous Delivery (CI/CD) pipeline for real-time integration. This will enable the model to continuously train on new data and provide real-time predictions. The model's ability to generalize effectively is supported by a wide range of data sources and ongoing learning through Federated Learning. The deep learning

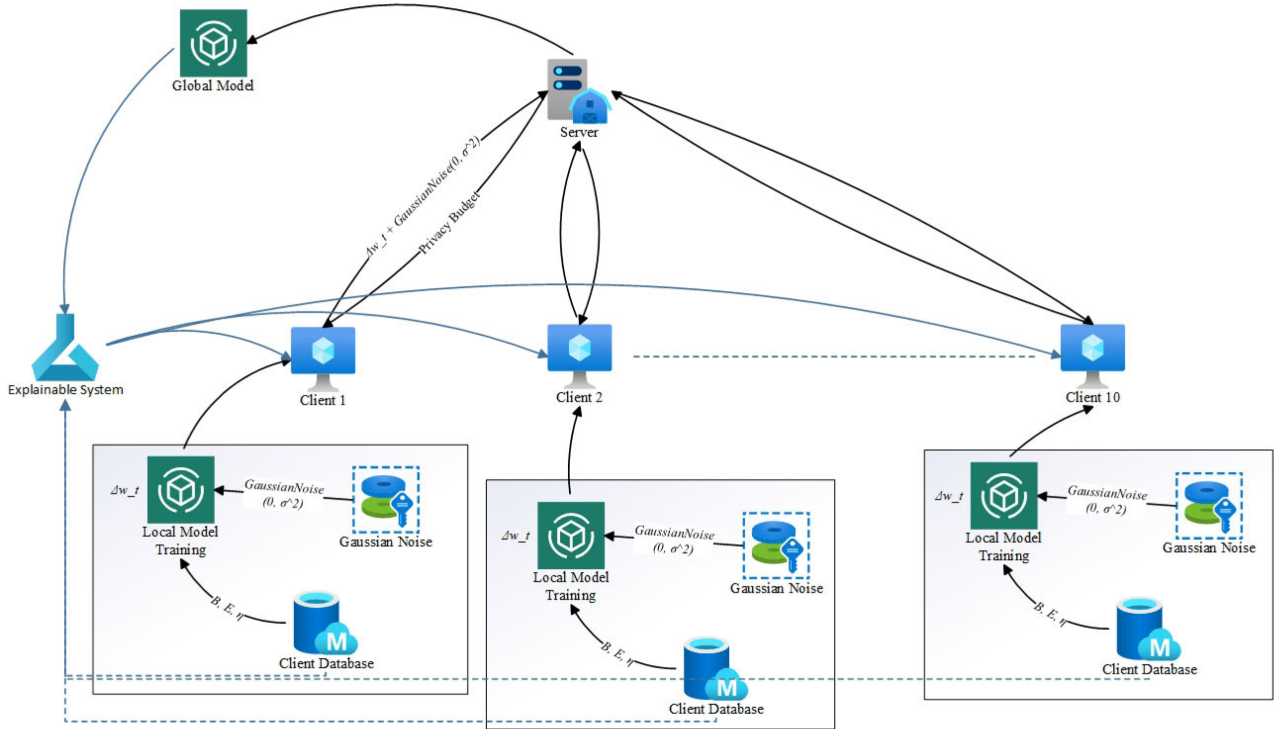

FIGURE 5. Proposed etup.

architectures excel at capturing intricate patterns, while Explainable AI ensures transparency in the model's decision-making process.

The fault detection model proposed in this study is currently tailored for the semiconductor wafer field, as it relies on and is trained with semiconductor wafer images. Nevertheless, the fundamental structure stated, which combines Federated Learning (FL), deep learning, and Explainable AI (XAI), exhibits notable flexibility and may be customized for different fault detection systems in the manufacturing industry and beyond. Industries that prioritize identifying faults or anomalies in their operations can derive significant advantages from the robust capabilities of this model. While the model was originally designed for semiconductors, its fundamental ideas can be applied to several domains, providing answers to sectors facing comparable issues of data privacy and the requirement for sophisticated pattern recognition.

The FL and XAI-based fault detection model is adaptable to evolving semiconductor processes, including nm-scale manufacturing, by leveraging continuous learning and robust deep learning capabilities. Successfully addressing challenges like increased data complexity and new defect types ensures the model's effectiveness in advanced technological environments.

The goal is to develop a reliable, private-preserving, and comprehensible model for semiconductor defect classification by combining FL with XAI. This method makes sure that the predictive potential of deep learning may be used while upholding strict guidelines for openness and result credibility.

IV. RESULTS AND DISCUSSION

This section explores the outcomes and analysis resulting from our research. We evaluate the performance of our models through a model evaluation, which involves analysing trends in training and generalisation. In addition, we offer insights into the internal mechanisms of our models using XAI visualisations, such as LIME and GradCAM. In addition, we consolidate the results obtained from the combination of FL and XAI, emphasising their joint influence on identifying faults in semiconductor design. By conducting this analysis, we provide a thorough comprehension of our study findings and their significance for the field.

A. MODEL PERFORMANCE EVALUATION

FL methods are investigated to improve semiconductor wafer defect detection accuracy. To achieve this goal, several complex deep learning models with various architectural advantages have been constructed and analyzed. The models studied were ResNet152, InceptionV3, DenseNet121, and MobileNetV2.

The evaluation process was intended to resemble FL's iterative nature. Ten rounds of training required each model to learn and adjust from a distributed dataset. This dataset simulates non-centralized data storage conditions. Training, validation, and test accuracy all served to evaluate model performance. These measures assess a model's learning, robustness, and generalization to unfamiliar data.

ResNet152 excelled in these arduous models. The model performed well with a peak training accuracy of 99.86%.

Its sophisticated depth and residual learning approach solves the vanishing gradient problem. Thus, the model learns from complex semiconductor wafer patterns. The model scored 98.63 % validation accuracy and 98.78% exceptional test accuracy. These data show the model's accuracy in defect detection and its ability to apply this talent to a variety of contexts, making it a good real-world choice.

The other contenders, however trailing ResNet152, performed well. InceptionV3's training accuracy was 98.61%, demonstrating its modules' efficiency in capturing cross-channel correlations. With its new dense connection structure, the DenseNet121 model achieved 99.72% training accuracy, proving feature reuse's effectiveness in learning. MobileNetV2, optimized for mobile and edge devices, achieved 98.82% training accuracy. Table 3 shows that lightweight deep learning models work in low-resource environments.

As models adjusted their parameters, training round accuracies converged. The models have hit their maximum data learning capacity, as projected. Given the random nature of neural network training, accuracy metrics varied within the expected range between rounds.

Validation and test accuracies have less fluctuation than training accuracies. These models' stability indicates their ability to generalize and makes them viable for situations where consistent performance on new data is more important than training. This is visible in fig. 6.

Comparing the work done by the authors in [40], which used a Deformable Convolutional Netwrok as referred to as the DCNet, which gave an accuracy of 93.20%, The architecture proposed here gave an exceptional result of 98.78% by using ResNet152 over 10 rounds of training distributed over 10 client data shards. The comparision between both of the works is compared in Table 5

B. TRAINING AND GENERALIZATION TRENDS

The training dynamics of deep learning models in semiconductor wafer defect detection revealed important insights into these complex systems' behavior during learning epochs. The tested models were hyperparameter optimized to find the best learning configurations. The model's hyperparameters were manually optimized through iterative adjustments based on experimental insights, enhancing its fault detection accuracy and adaptability in semiconductor manufacturing.

The models' performance improved significantly early in training, demonstrating a large gain in learning from the training data's wide range of attributes. During this stage of learning, models can extract and acquire the most important data features for defect detection.

After early training, accuracy improvement stagnated, approaching stability. This trend suggests that the models are performing at their best given their hyperparameter values and complexity. The accuracy plateauing occurs when the model's incremental learning advances get harder as it approaches its optimal state.

Each model's hyperparameters, the deep learning algorithm variables that control learning, were tuned. Opted hyperparameters included learning rate, batch size, epochs, and regularization terms. Model performance depended on learning rate. If set too high, the models would exceed the ideal state, while setting too low would impede convergence and waste computational resources. Higher batch sizes estimated the gradient more accurately but used more memory, affecting convergence stability. The number of epochs was chosen to provide models enough data to learn without overfitting. Regularization terms penalize model complexity, encouraging simpler and more generalizable data patterns and reducing overfitting.

Each model's training accuracy increased steadily, while its validation accuracy indicated how well it would perform on new data. Validation and training accuracies were nearly identical, with just minor differences. Close monitoring shows that the models learned generalizable patterns as well as training data.

Over 10 rounds, ResNet152's training accuracy increased, reaching a peak that shows its deep residual learning framework's effectiveness. The model's validation accuracy also increased, but little, due to its ability to perform well on data outside the training set. The accuracy gap between training and validation sets is ubiquitous in machine learning and often sought for model improvement.

The test accuracies, which determine model performance, remained consistent throughout the rounds. This stability means that the model can generalize well and is unaffected by unknown input, indicating that learning and generalization were successful. The hyperparameter setting is provided in Table 5.

C. EXPLAINABLE AI VISUAL INSIGHTS

By incorporating XAI into the model evaluation process, valuable insights were gained into the decision-making processes of our top-performing model, ResNet152, specifically in the domain of semiconductor wafer defect detection. This section explores the interpretive analysis enabled by several XAI methodologies, each revealing distinct aspects of the model's reasoning and attention focus.

1) GRADCAM HEATMAPS

The Grad-CAM technique expanded the investigation into the specific regions of interest identified by the model. Grad-CAM generated heatmaps by utilizing the gradients that enter the last convolutional layer of ResNet152. These heatmaps emphasized the areas that were most significant for the model's classifications. The heatmaps served as visual aids, exposing the spatial arrangement of focus inside the image. The locations with higher temperatures in these heatmaps corresponded to places that had a greater influence on the model's choice, providing a visually intuitive depiction of the model's attention. This analysis was particularly informative in verifying if the model's focus coincided with the known defect locations recognized by semiconductor specialists,

TABLE 4. Summary of the training results of federated learning over 10 rounds of training.

Round	*Model 1 - ResNet152*		*Model 2 - InceptionV3*		*Model 3 - DenseNet121*		*Model 4 - MobileNetV2*	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation
Round 1	98.24	97.99	96.73	95.45	97.35	97.11	97.51	97.29
Round 2	98.89	97.85	96.54	96.78	97.72	97.48	98.82	96.59
Round 3	98.45	98.14	96.65	96.21	98.35	98.12	96.34	96.65
Round 4	98.23	98.35	97.54	97.02	97.34	96.69	96.65	96.79
Round 5	97.76	98.33	96.48	95.97	98.31	96.91	96.87	96.91
Round 6	98.62	97.18	98.65	95.97	97.14	96.14	97.54	97.02
Round 7	99.58	98.61	98.66	96.11	98.15	96.77	99.3	97.46
Round 8	99.58	99.01	99.3	96.94	96.48	96.19	96.49	96.94
Round 9	99.86	98.69	99.37	96.94	98.35	99.33	99.37	97.48
Round 10	99.98	98.63	97.26	97.01	99.72	97.23	96.48	96.94
Test Accuracy	**98.78**		**96.94**		**97.61**		**96.81**	

TABLE 5. Comparison with one of the works.

Class	Proposed Work				Work done by authors in [41]		
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	Accuracy
C	0.99	1	1	0.99	0.94	0.91	0.99
C+EL	1	0.99	0.99	0.98	0.91	0.97	0.97
C+EL+L	0.99	0.98	0.99	0.97	0.95	0.93	0.96
C+EL+S	1	0.99	0.99	0.99	0.96	0.91	0.94
C+ER	1	1	1	0.99	0.93	0.97	0.99
C+ER+L	0.99	0.97	0.98	0.96	0.99	1	0.93
C+ER+S	0.99	1	0.99	0.99	0.9	0.94	0.95
C+L	0.99	0.98	0.99	0.97	0.6	0.88	0.93
C+L+EL+S	0.99	0.98	0.98	0.97	0.97	0.93	1
C+L+ER+S	0.96	0.99	0.97	0.98	0.94	0.94	0.99
C+L+S	0.98	0.99	0.99	0.98	0.92	0.99	0.97
C+S	0.99	1	0.99	0.98	0.92	0.96	0.98
D	0.99	1	1	0.99	0.97	0.89	0.96
D+EL	1	0.98	0.99	0.97	0.96	0.92	0.99
D+EL+L	1	0.99	0.99	0.97	0.91	0.98	0.96
D+EL+S	0.99	0.98	0.98	0.96	0.94	0.97	0.98
D+ER	0.99	0.99	0.99	0.97	0.96	0.94	0.93
D+ER+L	1	0.98	0.99	0.96	0.98	0.89	0.94
D+ER+S	0.97	0.99	0.98	0.97	0.94	0.91	0.92
D+L	1	0.98	0.99	0.96	0.95	0.91	0.94
D+L+EL+S	0.98	0.98	0.98	0.96	0.96	0.92	0.90
D+L+ER+S	0.96	0.99	0.98	0.97	0.98	0.88	0.90
D+L+S	0.98	0.99	0.99	0.96	0.99	0.96	0.88
D+S	0.99	1	1	0.97	0.92	1	0.89
EL	0.99	1	1	0.97	0.93	0.91	0.91
EL+L	0.99	0.97	0.98	0.94	0.97	0.97	0.92
EL+L+S	0.97	0.98	0.98	0.95	0.97	0.93	0.90
EL+S	1	0.99	0.99	0.97	0.95	0.91	0.88
ER	0.99	0.99	0.99	0.97	0.98	0.97	0.90
ER+L	0.99	0.96	0.98	0.94	0.89	1	0.92
ER+L+S	0.95	0.99	0.97	0.96	0.9	0.94	0.91
ER+S	0.99	1	0.99	0.97	0.99	0.88	0.88
L	0.99	0.98	0.99	0.95	0.97	0.93	0.86
L+S	0.98	0.99	0.99	0.96	0.98	0.94	0.89
N	1	1	1	0.97	0.96	0.99	0.87
NF	1	0.96	0.98	0.89	0.99	0.96	0.90
R	0.99	1	1	0.96	0.95	0.89	0.86
S	0.99	1	1	0.96	0.92	0.92	0.88

thus establishing the model's alignment with interpretability specific to the domain. The figures 7, 8, 9 show the GradCAM Heatmaps or the GradCAM Explanations of the original and the predicted classes.

2) LIME EXPLANATIONS

Through the process of simplifying the model's complicated decision-making process into a linear structure that is easier to grasp, LIME offered an alternative point of view. The contribution of specific pixels or segments in the image to the final prediction was discovered and quantified by LIME. This was accomplished by perturbing the input image and observing the changes that occurred in the model's output. In particular, distinguishing between characteristics that favorably or adversely influenced the categorization was made possible with the help of this pixel-wise breakdown, which was crucial in interpreting the rationale behind the model. These kinds of granular insights proved to

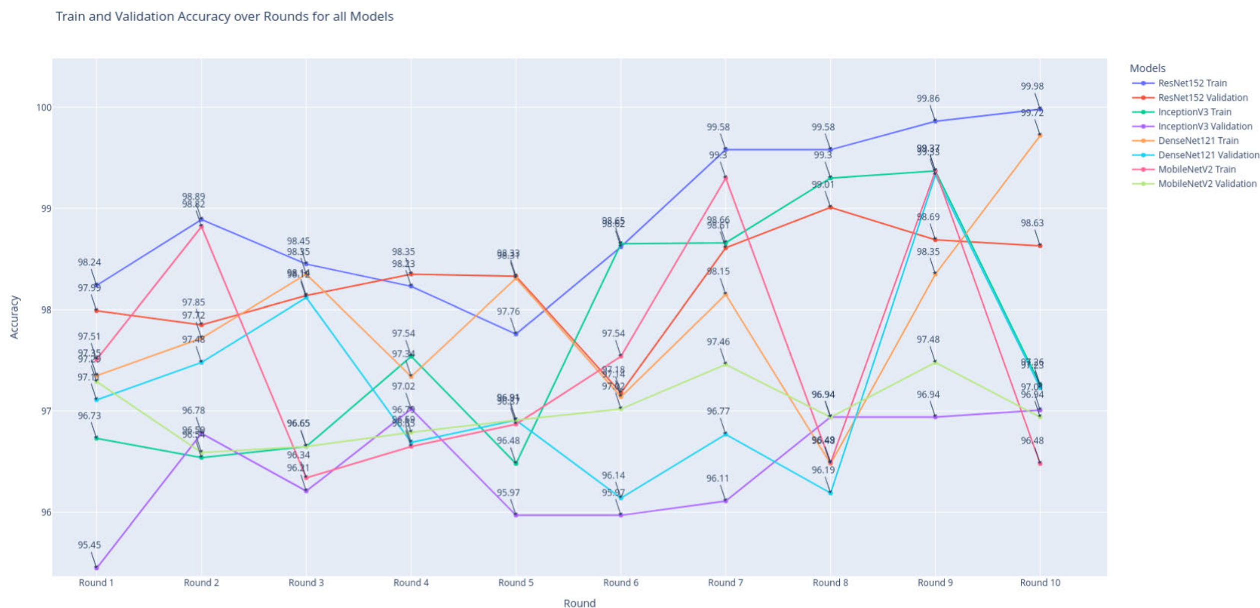

FIGURE 6. Training and Validation Plot of All the Models.

TABLE 6. Description of hyperparameters.

Hyperparameter	Description	Default Value
Learning Rate	The step size during optimization to reach a minimum.	0.001
Batch Size	Number of samples per gradient update.	32
Number of Epochs	Number of complete passes through the training dataset.	100
Optimizer	The method used to update weights in the neural network.	Adam
Momentum	Hyperparameter that accelerates SGD in the relevant direction and dampens oscillations.	0.9
Weight Decay (L2 Regularization)	A regularization technique adding a penalty for larger weights to the loss function.	0.0001
Dropout Rate	The fraction of the input units to drop to prevent overfitting during training.	0.5
Activation Function	The function used to introduce non-linearities into the network or to output the final result.	ReLU
Learning Rate Decay	The method used to reduce the learning rate over time.	None
Beta 1 (Adam Optimizer)	The exponential decay rate for the first moment estimates.	0.9
Beta 2 (Adam Optimizer)	The exponential decay rate for the second-moment estimates.	0.999
Epsilon (Adam Optimizer)	A small constant for numerical stability.	1e-8
Early Stopping	A form of regularization used to avoid overfitting by stopping the training process if the performance degrades on a held-out validation set.	False
Initialization Method	The method for initializing the weights in the network.	Glorot Uniform
Loss Function	The function used to compute the difference between the network's prediction and the actual label.	Categorical Crossentropy
Data Augmentation	Techniques used to increase the amount of data by adding slightly modified copies of already existing data.	False
Gradient Clipping	The method used to limit the size of gradients to prevent the exploding gradient problem.	False
Weight Constraint	Constraints that allow for specifying the norms of weights during optimization.	False
Learning Rate Scheduler	A technique used to adjust the learning rate during training.	False

be quite helpful in cross verifying the trustworthiness of the model and ensuring that it could concentrate on fault attributes that were truly significant, rather than being misled by noise or patterns that were irrelevant. LIME proved to give enhanced and effective explanations that are better as compared to all the Explainable AI methods that have been used, with a better pixel by pixel explanation score and super pixel highlights. The figures 10, 11, 12 show the LIME Explanation images for the said predicted classes.

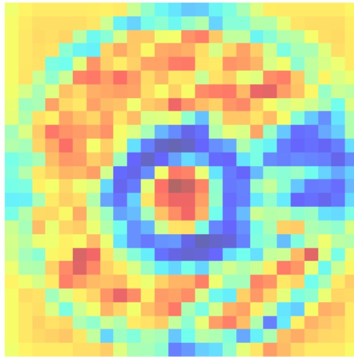

FIGURE 7. GradCAM Explanations for the Predicted Class: D+L+ER+S.

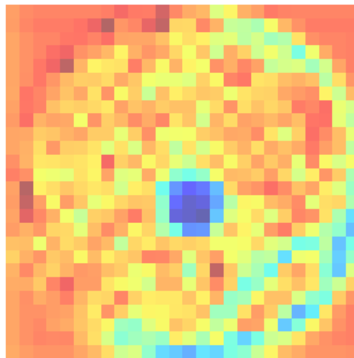

FIGURE 8. GradCAM Explanations for the Predicted Class: C+EL+S.

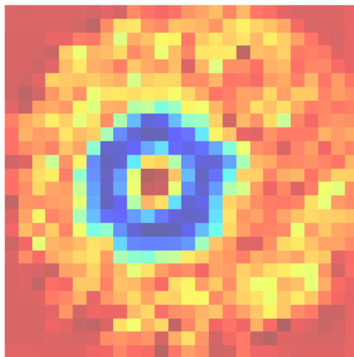

FIGURE 9. GradCAM Explanations for the Predicted Class: D.

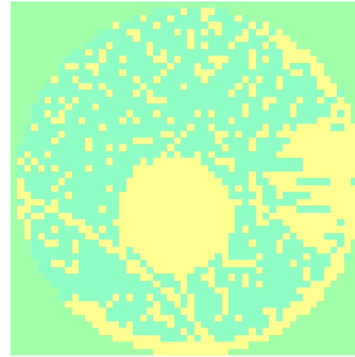

FIGURE 10. LIME Explanations for C+L+EL+S.

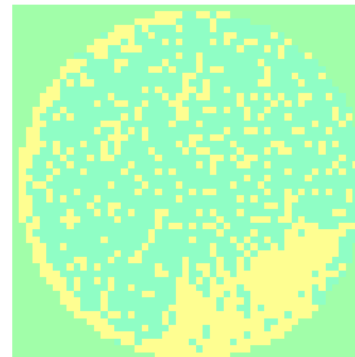

FIGURE 11. LIME Explanations for EL+L+S.

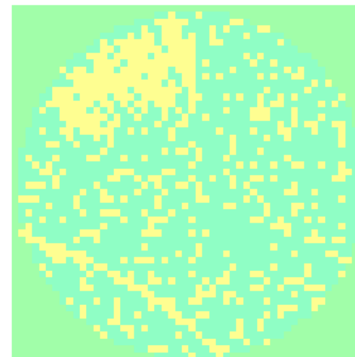

FIGURE 12. LIME Explanations for L+S.

D. SYNTHESIS OF FEDERATED LEARNING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

Integrating XAI methodologies with FL systems advances the search for transparent and reliable semiconductor wafer defect detecting AI solutions. This synthesis enhances the model's predictive power and interpretability, which is crucial in sensitive industries. The ResNet152 model, trained via FL and evaluated using multiple XAI methods, illustrates this integration.

The integration of XAI and FL yields advantageous outcomes. FL naturally addresses data privacy and decentralisation. This enables powerful model training without sensitive data. Combining this method with XAI secures remote data training and insight into model decision-making.

The ResNet152 model in a FL setting with remote nodes presented unique challenges and opportunities for XAI. The model learned new patterns and attributes throughout each training iteration utilizing different data subsets. A powerful XAI technique is needed to maintain interpretability and generalization in varied data distributions.

XAI visuals were tailored to FL. Grad-CAM heatmaps highlighted wafer imagery that was consistently selected as

relevant throughout training iterations and data points for a more complete view. LIME's local interpretable models showed how input changes affected model predictions, providing significant insights. This is crucial in federated situations with unpredictable data.

Applying XAI approaches to a FL model helped understand its learning dynamics. Despite training on scattered data, the findings confirmed the model's coherence in selecting defect-related attributes.

XAI in FL frameworks affects semiconductor production, notably wafer defect detection. It ensures that models are accurate, data protection-compliant, and transparent. Openness allows stakeholders to understand and verify AI-powered conclusions, boosting confidence. XAI techniques also help domain experts identify model biases and faults, enabling AI system improvement and fine-tuning.

V. FUTURE SCOPE AND DISCUSSIONS

The integration of FL and XAI into semiconductor failure detection is a manufacturing technology frontier with challenges and potential. Due to the high costs of model training and updates across geographically scattered nodes, FL systems require optimised communication methods. Communication efficiency affects the system's ability to recognise manufacturing faults and intervene quickly.

The development of algorithms that can navigate data heterogeneity between these nodes is also crucial. Such advances would ensure that model predictive performance is unaffected by data distribution, amount, and quality. These systems' capacity to scale large datasets from semiconductor manufacturing processes without compromising detection accuracy or computational efficiency is a promising research field.

Integration of these powerful AI technologies into semiconductor manufacturing infrastructure is difficult and requires novel solutions to assure interoperability and minimal operational impact. Quantum computing could improve data processing in these systems, enabling fault detection with remarkable precision and speed. FL and XAI in manufacturing require robust ethical and legal frameworks. These guidelines would promote ethical use and industry-wide adoption of these technologies.

Ultimately, interdisciplinary collaboration using semiconductor physics, materials science, and AI could accelerate defect detection system development. These systems would be more accurate, efficient, and able to provide deeper insights into the manufacturing process, leading the semiconductor industry to a future where defects are detected and preemptively addressed by AI.

VI. CONCLUSION

This paper has demonstrated that the integration of Federated Learning (FL) and Explainable AI (XAI) can achieve not only a high accuracy rate in fault detection within complex multi-stakeholder settings, such as those found in the semiconductor industry, but also ensure decision transparency and privacy protection. The application of XAI techniques has significantly improved the understanding of fault detection models developed through FL. By employing methods such as Grad-CAM heatmaps and LIME, we have gained deeper insights into how these models process data and make decisions. This has confirmed the models' ability to identify critical fault characteristics pertinent to semiconductor production and has ensured that, despite the distributed nature of the data and the learning process, the operations of the models remain transparent and intelligible. The synergy between FL and XAI techniques has proven crucial for maintaining trust in the models' predictions, especially in an industry where central sharing of sensitive design data is a significant concern. By achieving an exceptional level of accuracy in fault detection, as evidenced by the remarkable test accuracy of 98.78%, the proposed model sets the stage for significant enhancements in quality control measures. This could lead to a notable reduction in production faults and associated costs, marking a significant advancement towards more reliable and efficient semiconductor manufacturing processes. Additionally, the incorporation of XAI not only enhances the precision of these models but also adds a layer of transparency and accountability essential in sectors where the clarity and defensibility of decision-making processes are critical.

ACKNOWLEDGMENT

The authors would like to thank the "Centre of Next Generation Computing," Pandit Deendayal Energy University (PDEU), Gandhinagar. The support and resources provided by the Centre have been invaluable in the realization of this research. The Centre's state-of-the-art facilities and the dedicated assistance of its staff have significantly contributed to the advancements and findings presented in this article.

REFERENCES

[1] D. Pillai, "The future of semiconductor manufacturing," *IEEE Robot. Autom. Mag.*, vol. 13, no. 4, pp. 16–24, Dec. 2006.
[2] J. N. Burghartz, "Semiconductor manufacturing," in *Guide to State-of-the-Art Electron Devices*. Wiley, 2013.
[3] C. Brown and G. Linden, *Chips and Change: How Crisis Reshapes the Semiconductor Industry*. Cambridge, MA, USA: MIT Press, 2011.
[4] M. E. Cholette, M. Celen, D. Djurdjanovic, and J. D. Rasberry, "Condition monitoring and operational decision making in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 454–464, Nov. 2013.
[5] C.-H. Wang, W. Kuo, and H. Bensmail, "Detection and classification of defect patterns on semiconductor wafers," *IIE Trans.*, vol. 38, no. 12, pp. 1059–1068, Dec. 2006.
[6] M. Saqlain, Q. Abbas, and J. Y. Lee, "A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 3, pp. 436–444, Aug. 2020.
[7] J. Moyne and J. Iskandar, "Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing," *Processes*, vol. 5, no. 3, p. 39, Jul. 2017.
[8] S.-H. Huang and Y.-C. Pan, "Automated visual inspection in the semiconductor industry: A survey," *Comput. Ind.*, vol. 66, pp. 1–10, Jan. 2015.
[9] C. Constantinescu, "Trends and challenges in VLSI circuit reliability," *IEEE Micro*, vol. 23, no. 4, pp. 14–19, Jul. 2003.

[10] S. S. Fan, C.-Y. Hsu, D.-M. Tsai, F. He, and C.-C. Cheng, "Data-driven approach for fault detection and diagnostic in semiconductor manufacturing," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1925–1936, Oct. 2020.

[11] S. S. Fan, D.-M. Tsai, F. He, J.-Y. Huang, and C.-H. Jen, "Key parameter identification and defective wafer detection of semiconductor manufacturing processes using image processing techniques," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 4, pp. 544–552, Nov. 2019.

[12] C.-W. Liu and C.-F. Chien, "An intelligent system for wafer bin map defect diagnosis: An empirical study for semiconductor manufacturing," *Eng. Appl. Artif. Intell.*, vol. 26, nos. 5–6, pp. 1479–1486, May 2013.

[13] J. Wang, P. Gao, J. Zhang, C. Lu, and B. Shen, "Knowledge augmented broad learning system for computer vision based mixed-type defect detection in semiconductor manufacturing," *Robot. Comput.-Integr. Manuf.*, vol. 81, Jun. 2023, Art. no. 102513.

[14] G. Wen, Z. Gao, Q. Cai, Y. Wang, and S. Mei, "A novel method based on deep convolutional neural networks for wafer semiconductor surface defect inspection," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9668–9680, Dec. 2020.

[15] S. L. Yuen, C. W. Wong, P. Y. Lau, Z. Hussin, N. A. Kamarudin, M. H. Samsuri, M. S. M. Talib, and H. W. Hon, "GENSS: Defect classification method on extremely small datasets for semiconductor manufacturing," in *Proc. 27th Int. Comput. Sci. Eng. Conf. (ICSEC)*, Sep. 2023, pp. 419–424.

[16] F.-L. Chen and S.-F. Liu, "A neural-network approach to recognize defect spatial pattern in semiconductor fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 13, no. 3, pp. 366–373, Aug. 2000.

[17] B. M. Haddad, S. Yang, L. J. Karam, J. Ye, N. S. Patel, and M. W. Braun, "Multifeature, sparse-based approach for defects detection and classification in semiconductor units," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 145–159, Jan. 2018.

[18] P. B. Chou, A. R. Rao, M. C. Sturzenbecker, F. Y. Wu, and V. H. Brecher, "Automatic defect classification for semiconductor manufacturing," *Mach. Vis. Appl.*, vol. 9, no. 4, pp. 201–214, Feb. 1997.

[19] T. Evans, C. O. Retzlaff, C. Geißler, M. Kargl, M. Plass, H. Müller, T.-R. Kiehl, N. Zerbe, and A. Holzinger, "The explainability paradox: Challenges for xAI in digital pathology," *Future Gener. Comput. Syst.*, vol. 133, pp. 281–296, Aug. 2022.

[20] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller, Eds., "Interpretability in intelligent systems—A new concept?" *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Springer, 2019.

[21] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[22] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] P. Tchatchoua, G. Graton, M. Ouladsine, and J.-F. Christaud, "Application of 1D ResNet for multivariate fault detection on semiconductor manufacturing equipment," *Sensors*, vol. 23, no. 22, p. 9099, Nov. 2023.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[28] K. Limam, S. Cheema, S. Mouhoubi, and F. D. Freijedo, "Deep learning-based visual recognition for inline defects in production of semiconductors," *IEEE J. Emerg. Sel. Topics Ind. Electron.*, vol. 5, no. 1, pp. 203–211, Jan. 2024.

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[30] E. Shin and C. D. Yoo, "Efficient convolutional neural networks for semiconductor wafer bin map classification," *Sensors*, vol. 23, no. 4, p. 1926, Feb. 2023.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[32] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.

[33] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, A. Singh and J. Zhu, Eds., Apr. 2017, pp. 1273–1282.

[34] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[35] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecný, S. Mazzocchi, B. McMahan, and T. Van Overveldt, "Towards federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, vol. 1, A. Talwalkar, V. Smith, and M. Zaharia, Eds., 2019, pp. 374–388.

[36] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, Jan. 2019.

[37] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Oct. 2017. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2741

[38] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[40] J. Wang, C. Xu, Z. Yang, J. Zhang, and X. Li, "Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 4, pp. 587–596, Nov. 2020.

TANISH PATEL (Student Member, IEEE) is an Ambitious and Emerging Researcher with the Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, with a robust foundation in computer engineering, focusing on a broad array of interests include ML, CV, robotics, and cloud computing. His technical expertise is complemented by proficiency in multiple programming languages, alongside a deep understanding of machine learning and computer vision frameworks like TensorFlow and OpenCV. He has actively engaged in several innovative projects. He is also an AI Engineer with Paperchase Inc., New York City. He is also the Founder and the Director of TechRim InfoTech Private Ltd. His dedication to technological innovation is also evident in his participation in various technical symposiums and hackathons and his leadership roles in student technical organizations. His work embodies a commitment to leveraging cutting-edge technology to address real-world challenges, making significant strides toward the advancement of efficient and sustainable technological solutions. His commitment to technological innovation is matched by his dedication to community building and skill development within his peer networks.

RAMALINGAM MURUGAN is currently a Senior Assistant Professor with the School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India. He has completed his research in the field of vehicular ad-hoc networks, in 2020. His research interests include federated learning, machine learning, computer vision, the Internet of Things, deep neural networks, blockchain, and generative AI.

GOKUL YENDURI received the M.Tech. degree in IT from Vellore Institute of Technology, in 2013, where he is currently pursuing the Ph.D. degree. He is also an Assistant Professor with VIT-AP University. He is also a Senior Research Fellow with the DIVERSASIA Project, co-funded by the Erasmus+ Program of European Union in the past. He has attended several national and international conferences, workshops, and guest lectures; and has published articles in peer-reviewed international journals. His research interests include machine learning and predictive analysis, software engineering, assistive technologies, and metaverse. He is acting as a reviewer of many prestigious peer-reviewed international journals.

RUTVIJ H. JHAVERI (Senior Member, IEEE) received the Ph.D. degree in computer engineering, in 2016. He is currently an Experienced Educator and a Researcher with the Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, India. He conducted his postdoctoral research with Delta-NTU Corporate Laboratory for Cyber-Physical Systems, Nanyang Technological University, Singapore. In 2017, he was awarded with prestigious Pedagogical Innovation Award by Gujarat Technological University. He is also co-investigating a funded project from GUJCOST. He was ranked among top 2% scientists around the world, in 2023, 2022, and 2021. He has 4000 Google Scholar citations with H-index 35. He is an Editorial Board Member in various journals of repute, including IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS and *Scientific Reports*. He also serves as a reviewer for several international journals and also as an advisory/TPC member in renowned international conferences. He has authored more than 160 articles, including IEEE/ACM TRANSACTIONS and flagship IEEE/ACM conferences. Moreover, he has several national and international patents and copyrights to his name. He also possesses memberships of various technical bodies, such as ACM, CSI, and ISTE. He is the Coordinator of SCAN—Smart Cities Air Quality Network. Moreover, he is a member of the Advisory Board in Symbiosis Institute of Digital and Telecom Management and other reputed universities, since 2022. He is an editorial board member in several Springer and Hindawi journals. He also served as a Committee Member for "Smart Village Project," Government of Gujarat, at the district level, in 2017. His research interests include cyber security, the IoT systems, SDN, and smart healthcare.

HICHEM SNOUSSI received the Diploma degree in electrical engineering from the Ecole Superieure d'Electricite (Supelec), Gif-sur-Yvette, France, in 2000, the D.E.A. and Ph.D. degrees in signal processing from the University of Paris-Sud, Orsay, France, in 2000 and 2003, respectively, and the H.D.R. degree from the University of Technology of Compiegne, in 2009. Between 2003 and 2004, he was a Postdoctoral Researcher with IRC-CyN, Institut de Recherches en Communications et Cybernetiques de Nantes. He has spent short periods as a Visiting Scientist with the Brain Science Institute, RIKEN, Japan, and Olin Neuropsychiatry Research Center, Institute of Living, USA. Between 2005 and 2009, he was an Associate Professor with the University of Technology of Troyes, France, where he has been a Full Professor, since 2010. His research interests include Bayesian techniques for source separation, information geometry, differential geometry, and machine learning. Since 2010, he has been in charge of the CapSec Plateform (Sensors for Security). He is the principal investigator of many research projects and industrial partnerships. Between 2016 and 2020, he has been the Manager of the LM2S Laboratory, Charles Delaunay Institute. Since 2021, he has been the Deputy Director of the LIST3N Laboratory and the Team Leader of the M2S Group. He co-founded two start-ups: AQUILAE (computer vision), in 2017, and Damavan Imaging (nuclear Compton cameras), in 2014.

TAREK GABER (Member, IEEE) received the Ph.D. degree in computer science, with a focus on information security from The University of Manchester, in 2012. He is currently a Senior Lecturer and the Programme Leader for the M.Sc. in Cyber Security, School of Science, Engineering & Environment, University of Salford, Manchester, U.K. He has held positions at several universities, including The University of Manchester, U.K.; Suez Canal University, Egypt; and VSB Technical University of Ostrava, Czech Republic. Over the past three years, he has secured a total of 600 in funding to support his research from various public funding bodies, including Innovate U.K., GCHQ, and UKAEA. This funding has facilitated the development of software tools (in AI and cybersecurity) based on his published articles. He has served as a keynote speaker and the co-chair at several international conferences. He has also acted as a Lead Guest Editor for many SCI-indexed international journals, including the *Journal of Healthcare Engineering*, *Applied Sciences*, *Wireless Communications and Mobile Computing*, *Sustainability*, and *Electronics*. With over 100 publications in international journals, conferences, and book chapters. His primary research interests include cybersecurity, machine learning, artificial intelligence, and secure software engineering.