Mixture-of-Experts for Multi-Domain Defect Identification in Non-Destructive Inspection

Venkata Devesh Reddy Seethi*, Ashiqur Rahman*, Austin Yunker†, Rami Lake*, Zachary Kral‡, Rajkumar Kettimuthu†, and Hamed Alhoori*
* Department of Computer Science, Northern Illinois University, DeKalb, IL, USA
† Argonne National Laboratory, Lemont, IL, USA
‡ Spirit AeroSystems, Wichita, KS, USA

***Abstract*—Composite materials are widely used in aircraft structures because of their superior mechanical properties. However, their complex failure modes require sophisticated inspection methods to ensure structural integrity. Ultrasonic testing (UT) is a common non-destructive inspection (NDI) technique for aircraft composites that can detect internal and external defects with high resolution and accuracy. Despite their effectiveness, traditional UT methods rely on the manual interpretation of ultrasonic signals, which is time-consuming, labor-intensive, and subjective. Furthermore, processing such large-scale data, particularly across materials of varying thicknesses, significantly increases the computational demands of deep learning model optimization. To overcome these challenges, we propose an efficient sparse mixture-of-experts (MoE) model with a multi-level loss function and introduce four novel training objectives to improve computational efficiency and accuracy in identifying surface defects in composite aircraft materials. We evaluated our approach on material with multiple thicknesses or domains comprising various defects. Our experimental results demonstrate higher accuracy and F1-Score, with only 10% training epochs compared to baseline MoE.**

***Index Terms*—Mixture of experts, efficient machine learning, non destructive inspection, composite materials, ultrasonic signals**

I. Introduction

The introduction of NDI 4.0 into aircraft fuselage inspection, utilizing AI and cyber-physical systems [1], represents a crucial advancement in non-destructive inspection. The CFRPs have become leading materials in manufacturing due to their superior fatigue tolerance, mechanical durability, lower carbon footprint, and improving fuel efficiency of vehicles [19]. While CFRP offers numerous advantages, its complex nature demands a thorough inspection from highly skilled individuals to promptly identify and assess structural failures throughout their lifecycle to ensure safety and minimize costs [3]. The structural inspection needs of CFRP materials are met through various sensor modalities [6], with ultrasonic testing (UT) serving as the industry standard for NDI, specifically in aircraft manufacturing. The first advantage of ultrasonic signals is their ability to *differentiate* normal and defective regions through attenuation of signals [9]. Second, its high-resolution imaging capability enables *fine-grained* detection of internal and external defects in composite materials with high precision. However, conventional UT methods require manual interpretation of the ultrasonic signals, which is time-consuming and labor-intensive [20]. Therefore, designing effective AI systems is essential to alleviate the manual workload on inspectors and steer progress towards zero-defect manufacturing [11].

While domain expertise in NDI allows inspectors to locate defects effectively, the sheer volume of data generated - often comprising millions of pixels places a considerable cognitive burden on the inspectors [10]. To alleviate this burden and tedious elements of manual inspection, the future of NDI lies in reshaping the process with AI systems designed to be human-in-the-loop [1]. By melding human expertise with AI's analytical capabilities, a human-in-the-loop framework can transform aircraft fuselage inspections during production with precise detection of defects in complex materials such as carbon fiber-reinforced polymer composites (CFRP). Numerous studies have shown promising outcomes in such AI-assisted UT both in general applications [21, 16] and inspection of aircraft fuselages [22, 15]. Designing AI systems for defect identification in NDI has several challenges: **big-data** due to the immense volume of data generated by scans, often comprising millions of pixels and handling data from **multiple domains** (i.e., varying material thicknesses). These challenges significantly increase the complexity and resource requirements for training deep learning models. To address these issues, we present a sparse mixture-of-experts (MoE) architecture for defect classification in multi-domain data used in the aerospace industry. Additionally, we formulate a novel multi-level objective with four new custom loss functions. Empirical results show our proposed method has improved performance and efficiency over the baseline MoE model.

II. Related Works

AI-driven non-destructive inspection (NDI) has demonstrated success in different testing modalities in the industry, such as acoustic emission testing, ultrasonic testing (UT), eddy current testing, and X-ray computed tomography [3, 6]. In civil applications, AI systems have been used for structural health monitoring of dam slopes using UT [16]. Within the aerospace domain, AI systems evolved from a heuristic-based approach to automated AI systems, marking a significant development, as demonstrated by Kral et al. [9]. Kokurov et al. [8] investigated ultrasonic methods for detecting defects in laminated composite materials, concluding their effectiveness. Bettayeb et al. [2] demonstrated the efficacy of discrete

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00200

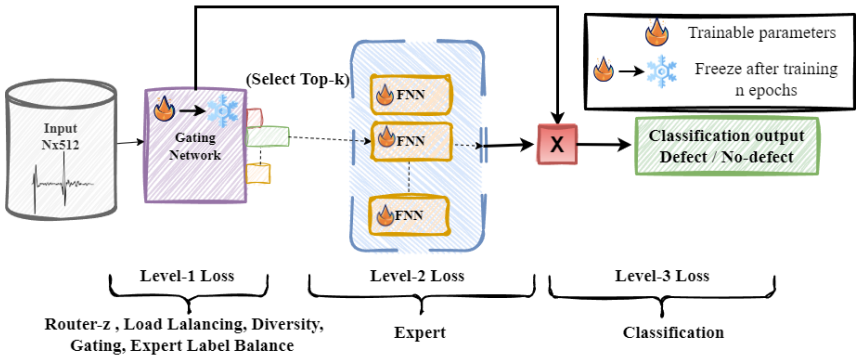

Fig. 1. Sparse Mixture of Experts (MoE) architecture for ultrasonic waveform processing. The input waveform is routed to a specific expert (a feedforward neural network) based on router probabilities, which are multiplied with the selected expert's output, producing the final result.

wavelet transform in defect classification. Meng et al. [14] developed a deep learning framework for classifying ultrasonic signals from CFRP specimens using wavelet transform for preprocessing. McKnight et al. [13] explored alternative AI applications in synthetic data generation methods for domain adaptation in ultrasonic signals. AI systems have also become a cornerstone for the predictive maintenance of aircraft fuselages. Studies such as Prakash et al. [15] and Ye et al. [21] have utilized manual UT phased array data from Glass Reinforced (GLARE) FML of A380 aircraft and conducted comprehensive NDI on steel plates with different types of flaws. Kral [9] proposed an ANN network for detecting defects in aluminum sheets and aircraft fuselage. Prakash et al. [15] proposed a defect detection framework using a histogram of oriented gradients as preprocessing and support vector machines on GLARE FML data from the A380. Yunker et al. [22] adopted a 3D-UNet for identifying defects in a 3D scan data from an industry-manufactured aircraft. These studies validate the effectiveness of AI in defect identification. Therefore, we address the question: *how can we make AI models efficient for NDI for multi-domain materials?*

The MoE architecture is a plausible solution for NDI to address the computational needs and multi-domain data. Research in the MoE paradigm has been active for nearly 30+ years, where recent works showcase their desirable properties in terms of better computation efficiency, improved data sampling, and higher accuracy [4]. Moreover, works in MoE have also achieved commendable results for domain adaptation where the criteria are to generalize and adapt over new unseen domains [12] and domain-aware models [7]. A mixture of Encoder-Decoder models such as BLIP has also successfully handled multi-modal data for automotive quality inspection [18]. Despite these advancements, MoE remains underexplored for NDI applications. We aim to address this gap with our work and propose novel training objectives that *reduce* computational overload and *improve* performance.

III. Dataset

Domain Information: While the surface of a manufactured part initially goes through visual inspection, that is not adequate to detect subsurface defects. To identify subsurface defects in composites, non-destructive inspection (NDI) methods are utilized. The primary inspection method is through ultrasonic testing, which sends a beat/impulse strain wave with a frequency in the ultrasonic range through a material. This method is not limited by the material type, as the signal travels in a non-linear fashion and reflects/disperses from every different material boundary, geometry wall, or defect region. Our goal in this study is to build scalable models that aid in decreasing the evaluation effort through signal processing. We use **standard scans** in our study that are test specimens and not actual airplane parts. Standard scans are equivalent references of the real aircraft fuselage, which are defined by the owner of the final product. Figure 2 shows a visual example of the standard scan with annotated defects across different thickness levels categorized as domains 1-7.

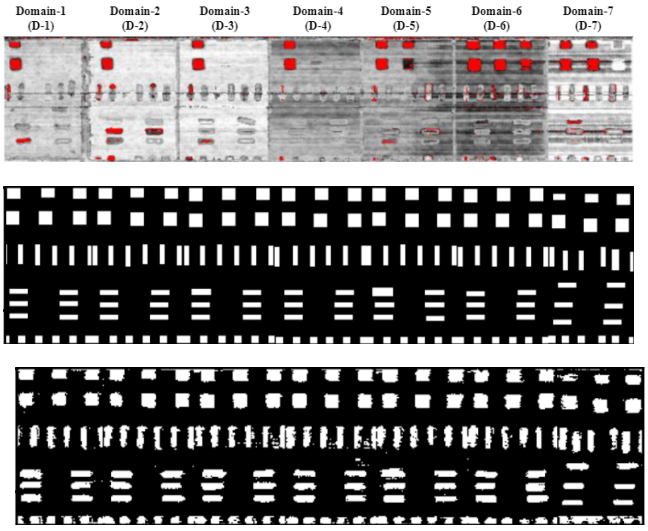

Fig. 2. **Top:** One standard scan of dimensions $[128, 533]$ visualized in $2D$ on, with thresholding scheme from software in red. The scan is split into seven domains, labeled $D1$-7, based on material thickness. **Middle:** The ground truth (i.e., annotated defects), where the white region is a defect. **Bottom:** the predictions from our model.

Dataset Description and Preprocessing: Preprocessing the ultrasonic signals is important to mitigate noise and ensure consistency across the dataset. Therefore, we first apply a Hilbert transform to eliminate high-frequency noise, followed by min-max normalization to avoid bias towards either of the peaks at $+/-20000$. To further standardize the signals, we align the peaks caused by variations in the scanner apparatus. Specifically, we identify the first peak in the signal and shift it to a fixed point in the spectrum. Yet another common practice is to incorporate discrete Fourier transform. However, this would not be helpful as the Fourier transforms do not preserve the signal structure, which makes it harder to interpret visually. In Figure 3, we visualize the preprocessed signal of a random defect point and a random non-defect point from Domain-1 ($D-1$) from the scan shown in Figure 2.

Our dataset contained 31 scans, split into 11 for training, 11 for validation, and 9 for testing. Each scan has height, width, and depth dimensions of 128, 533, and 512, respectively. Therefore, each of our training and validation data is composed of $128 \times 533 \times 11 = 750464$ ultrasonic signals and have dataset dimensions of $[750464, 512]$ and test dataset of dimensions $[614016, 512]$. During training, we balance the batches using random weighted sampling across the domain and the defect/ non-defect classes.

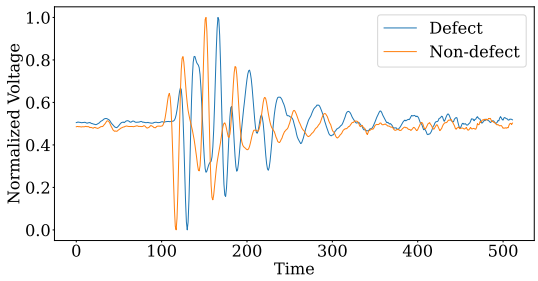

Fig. 3. Ultrasonic signals from a defect and a non-defect region from D-1.

IV. Methods

We present our architecture in Figure 1, which consists of a gating network, multiple feedforward networks acting as experts, multiple training objectives, and a sparse routing system. In this section, we detail each component.

A. Sparse Mixture-of-Experts

The Sparse MoE model follows an ensembling architecture that combines the strengths of multiple expert models to achieve better sample efficiency and improve overall performance. In this model, each input is routed to a subset of experts through the gating network, and the outputs from each selected expert are combined to form the final output. The gating network generates weights as softmax probabilities based on gating logits of size N, where N is the number of experts in our network. The gating function $G(x)$ for an input x is expressed as $G(x) = Softmax(g_i(x))$. Here, $g(x)$ is the gating logits vector of size N, corresponding to the expert outputs. While the training objectives use $g(x)$ to compute auxiliary loss, the selection criteria for experts and the weighting factor are based on $G(x)$. The sparse MoE also uses a sparse routing scheme [17] where only top-k experts are chosen. In the case of top-1 routing, only one expert is chosen. Additionally, we freeze the gating network while using our proposed training objectives without any degradation in model performance. This reduces the requirement of gradient adjustments to the gating network, reducing the computation load. Our empirical investigation shows that the gating network can be frozen after 10 epochs.

B. Training Objectives

Previous works in MoE have extensively explored various auxiliary training objectives to optimize models for efficiency, performance, or both. Shazeer et al. [17] proposed the idea of load balancing for balanced expert usage. Following this Zoph et al. [23] proposed a simple differentiable load balancing loss function, which we adopt it in our approach as a baseline. Additionally, some works also proposed domain-specific loss [7] for building domain-aware models. Following these works, we craft our training losses to optimize for classification with domain awareness and stability.

We framed multi-level losses for our model to improve the sample efficiency and performance after exploring several training losses from the existing literature. We divide our losses in levels 1, 2, and 3, as shown in Figure 1. At **level-**3, we employ classification loss ($\mathcal{L}_{\text{C}}$) which operates on the final predictions. The $\mathcal{L}_{\text{C}}$ jointly optimizes the collective performance of all the experts and the gating function. At **level-**2, we introduce an expert-specific loss, ($\mathcal{L}_{\text{E}}$), that operates at expert level. Finally, at **level-**1, we address inefficient routing and mode collapse with Load Balancing Loss ($\mathcal{L}_{\text{LB}}$) and Router-z-loss ($\mathcal{L}_{\text{RZ}}$) for efficient routing. Additionally, we propose new loss functions such as diversity loss ($\mathcal{L}_{\text{D}}$) and expert label balance loss ($\mathcal{L}_{\text{ELB}}$), which we describe below.

Classification Loss ($\mathcal{L}_{\text{C}}$): The classification loss is a binary cross-entropy (BCE) loss objective that computes the error in classifying normal and defect signals. This can be simply represented as $\mathcal{L}_{\text{C}} = BCE(y_i, \hat{y}_i)$, where $\hat{y}_i$,y$\in \{0, 1\}$ is the predicted probability and ground truth, respectively. The binary cross-entropy loss encourages the entire model to output probabilities close to the true labels.

Expert Loss ($\mathcal{L}_{\text{E}}$): Unlike $\mathcal{L}_{\text{C}}$ which works on a global level, the expert loss focuses on optimizing each individual expert. This loss assesses how well each expert performs on the specific data routed to it, enabling more fine-grained, targeted optimization. The expert loss is formulated similarly to classification loss, as shown below.

$$\mathcal{L}_{\text{E}} = -\frac{1}{N}\sum_{i}^{N}\frac{1}{M_i}\sum_{j}^{M_i} y_j^{(i)}\log(\hat{y}_j^{(i)}) + (1-y_j^{(i)})\log(1-\hat{y}_j^{(i)})$$

In this expression, $y_j^{(i)}$ denotes the true label of the i-th sample processed by expert j, and $\hat{y}_j^{(i)}$ represents the predicted probability output by expert j for the i-th sample. The summation is performed over all experts $i \in N$ within their respective sample subsets M_i.

Load Balancing Loss ($\mathcal{L}_{\text{LB}}$): One challenge in training an MoE model is to ensure the balanced contribution of all experts. Without proper balance, some experts may dominate the workload, leading to inefficiency. To mitigate this challenge, we adopt the load balancing loss function similar to that used in switch transformers [5].

$$\mathcal{L}_{\text{LB}} = N \sum_{i=1}^{k} P_i(x) \cdot f_i(x)$$

where f_j is the fraction samples routed to expert i, and P_j is the fraction of the router probability allocated for expert j,

$$f_i = \frac{1}{M_i} \sum_{x \in M_i} \mathbb{1}\{\text{argmax} g(x), i\}, P_i = \frac{1}{M_i} \sum_{x \in M_i} G_i(x)$$

where m_i denotes a batch within M, such that $m_i \in M$, and $G_i(x)$ is the gating probability for expert i given input x.

Router-z Loss ($\mathcal{L}_{\text{RZ}}$): The gating layer is susceptible to large logits which hinders convergence to optimal routing. Therefore, we employ router z-loss, inspired by the work of Zoph et al. [23]. We formulate router z-loss below as $\mathcal{L}_{\text{RZ}}$. Here, N is the number of experts, M is the number of data points entering the experts, and $g_j^{(i)}$ are the logits from the gating layer.

$$\mathcal{L}_{\text{RZ}} = \frac{1}{M} \sum_{i=1}^{M} \left(\log \sum_{j=1}^{N} e^{g_j^{(i)}} \right)^2$$

Gating Loss ($\mathcal{L}_{\text{G}}$): We design gating loss as a cross-entropy (CE) loss similar to Jain et al. [7] where the true labels are domain labels, d and prediction is the softmax output of gating network G. Here, $G(x_i^{(j)})$ is the softmax output of the gating layer for input sample x_i. Where $j \in M_i$ are samples going through individual expert i and $i \in D$ is the number of domains as shown in Figure 2. The gating loss works only when the number of experts is equal to the domains. Hence, we set $N = D$.

$$\mathcal{L}_{\text{G}} = -\sum_{i=1}^{D} \frac{1}{M_i} \sum_{j=1}^{M_i} log(G(x_i^{(j)}))$$

Diversity Loss ($\mathcal{L}_D$): We define the diversity loss as $-\sum \text{Var}(g_j)$, which is the summation of the variance of all gating logits for *N* experts in our model. The negative sign indicates that higher diversity is rewarded. This loss function encourages the experts to have diverse output distributions, prompting better specialization.

Expert Label Balance Loss ($\mathcal{L}_{ELB}$): We formulate expert label balance loss as:

$$\mathcal{L}_{ELB} = \frac{1}{N} \sum_{i=1}^{N} (R_j^{(i)} - S_j^{(i)})^2$$

where N is the number of experts and M is the number of samples, $R^{(i)} = \sum_{k \in M}^{R} y_k$ is the sum of the defect signals routed to the i-th expert, and $S^{(i)} = \sum_{k \in M}^{S} y_k$ is the sum of the non-defect signals routed for the i-th expert. The mean of the squared difference between the sum of positive samples, R_i, and the sum of negative samples, S_i. This loss encourages the gating network to route a balanced number of defect and non-defect samples to each expert.

Total Loss ($\mathcal{L}_{\text{total}}$): Our proposed model utilizes all the proposed loss functions in the form of a weighted sum. Specifically, we define our total loss as:

$$\mathcal{L}_{\text{total}} = \alpha_C \cdot \mathcal{L}_{\text{C}} + \alpha_G \cdot \mathcal{L}_G + \alpha_{LB} \cdot \mathcal{L}_{LB} + \alpha_{RZ} \cdot \mathcal{L}_{RZ} + \alpha_E \cdot \mathcal{L}_E + \alpha_D \cdot \mathcal{L}_D + \alpha_{ELB} \cdot \mathcal{L}_{ELB}$$

where $\mathcal{L}_C$, $\mathcal{L}_G$, $\mathcal{L}_{LB}$, $\mathcal{L}_{RZ}$, $\mathcal{L}_E$, $\mathcal{L}_D$, and $\mathcal{L}_{ELB}$ represent the classification loss, gating loss, load balancing loss, router z-loss, expert loss, diversity loss, and expert label balance loss, respectively. The α terms represent the weights assigned to each loss term.

C. Model Hyper-parameters

We have fixed the number of experts as 7 to account for the 7 domains in our dataset. Next, we set the hidden layers in the gating network and each expert as 2 and 4, respectively, after doing a layer sweep from 1-10 layers. By fixing the model architecture, we isolate and assess the impact of our proposed loss functions. Furthermore, to maintain network homogeneity, we use hidden-layer dimension of 1024. Finally, we tested different values of α_i through a randomized search between $1e^{-3}$ to 1 on a logarithmic scale. The value of α_i controls the contribution of each loss term in the total loss. We set α_C, α_G, α_E as 1, α_D, α_{ELB} as $1e^{-2}$, α_{RZ}, α_{LB} as $1e^{-1}$.

V. Results

A. Metrics

To account for the class imbalance in our dataset, we selected metrics that appropriately weigh both the defect and non-defect classes. First, we use balanced accuracy BA that averages the defect and non-defect class accuracies as $BA = (A_D + A_{ND})/2$, where A_D and A_{ND} are accuracies of the defect and non-defect classes, respectively. Next, we use F1-Score (F1) with macro averaging, which, similar to balanced accuracy, averages scores for defect and non-defect classes: $F1 = (F1_D + F1_{ND})/2$ where $F1_D$ and $F1_{ND}$ are F1-Scores for defect and non-defect classes, respectively. The F1-Score is computed as: $\text{F1} = (2 * \text{P} * \text{R})/(\text{P} + \text{R})$. Where recall is $\text{R} = \text{TP}/(\text{TP} + \text{FN})$, and precision is $\text{P} = \text{TP}/(\text{TP} + \text{FP})$. Here TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively.

TABLE I
EVALUATION OF OUR PROPOSED MODEL WITH BASELINE. LOWER IS BETTER FOR CONVERGENCE AND HIGHER IS BETTER FOR PERFORMANCE METRICS.

Architecture	**Loss functions**	**Convergence↓**		**Performance Metrics↑**	
		Checkpoint-1	**Checkpoint-2**	**F1-Score**	**Accuracy**
Top-7 MoE	$\mathcal{L}_C$	91	322	0.846	89.6%
Top-1 MoE	$\mathcal{L}_C$	119	253	0.846	89.6%
(Baseline) Top-1 MoE	Standard losses	321	398	0.853	89.4%
(Ours) Top-1 MoE	Proposed losses	43	288	**0.869**	**91.3%**
(Ours) Top-1 MoE	Standard + proposed losses	**33**	**205**	0.859	90.8%

Standard losses: $\mathcal{L}_C$ + $\mathcal{L}_{LB}$ + $\mathcal{L}_{RZ}$, **Proposed losses:** $\mathcal{L}_G$ + $\mathcal{L}_E$ + $\mathcal{L}_D$ + $\mathcal{L}_{ELB}$
Interpretation: Lesser is better for convergence, and more is better for performance metrics
Checkpoint-1 (C-1): Epoch where validation F1-Score is 0.85 (early stopping)
Checkpoint-2 (C-2): Epoch where validation F1-Score is *maximum* in 400 epochs

B. Experimental Setup

Our architecture leverages the strengths of multiple experts, each specializing in different aspects of defect classification. In addition to the standard training objectives, we propose new objectives that show improved performance and effective utilization of training data and expert networks for faster convergence. Our criteria for benchmarking are based on ① performance metrics and ② convergence, providing insights into both effectiveness and efficiency, respectively. For evaluation, we use balanced accuracy and macro F1 to test our models and report results in Table I. We also perform an ablation for our proposed training objectives in Table II.

Performance Analysis. To assess ① performance by first comparing top-7 MoE to top-1 MoE where they both achieve similar performance; however, top-7 MoE has a larger computation requirement. We discuss the importance of top-1 routing in more detail in Section V-C. Next, we compare our proposed loss functions with two baselines: top-1 MoE with only classification loss (i.e., L_C) and top-1 MoE with standard losses (i.e., L_C, L_{BL} and L_{ZL}). In both cases, we do not see any improvement in performance, however, the standard losses require more epochs to converge. Although we do not see a major improvement in performance (i.e., F1-Score and accuracy) compared to baseline models, there is a significant improvement in convergence. Our method achieves an F1-Score of 0.869 and an accuracy of 91.2%, whereas the baseline model achieves an F1-Score and accuracy of 0.853 and $89.\%$, respectively. These results further underscore the effectiveness of our approach in accurately classifying defects in ultrasonic data.

Turning to assess the ② convergence of our models, we set two checkpoints, checkpoints-1 and 2. The **Checkpoint-**1 is the epoch where validation F1-Score is 0.85. **Checkpoint-**2 **(C-**2**):** on the other hand is the epoch where validation F1-Score is *maximum* among the 400 epochs. Our proposed model reaches checkpoint-1 in 33 epochs compared to 321 epochs in the baseline model using standard loss functions, marking a significant improvement. Furthermore, even without standard loss functions and just adopting our loss functions, the model reaches checkpoint-1 in 43 epochs. We also see an improvement in the checkpoint-2 where our model requires 208 epochs compared to 398 for the baseline model. Finally, the average time taken to train per epoch for the top-7, top-1 MoE with only $\mathcal{L}_C$ was 12.2 and 6.2 seconds, respectively. For top-1 MoE with only standard losses or only proposed losses, the model takes 6.3 seconds. When combining both standard and proposed losses, it increases to 6.5 seconds. These results demonstrate the efficiency of our proposed training objectives in accelerating convergence.

C. Ablation Study

To ensure a fair ablation study, we keep the parameters across all models consistent and focus on checkpoint-1 - with early stopping criteria on validation F1-Score of 0.85 instead of checkpoint-2 which is the epoch where validation accuracy is maximum.

Importance of Top-k Routing Mechanism. We investigate the importance of the routing mechanism in the rows $1-2$ of Table I. Firstly, we implement an MoE model with a soft routing mechanism with top-7 routing where all experts are always active and contribute to the output with different weights. A more efficient alternative is to use top-1 routing, where only one expert is active. For instance, consider our MoE architecture where the gating network has 2 hidden layers (1.584M parameters for the gating layer) and each of 7 experts has 4 hidden layers (i.e., 3.683M Parameters for each expert). In top-7 gating approach, the model would utilize the entire model's parameters, which is $1.584 + 3.683 * 7 = 27.365M$ parameters. However, using the top-1 routing, the model only uses $1.584 + 3.683 * 1 = 5.267M$ - about 1/5th of the entire model's parameters.

Importance of Standard Loss Functions. Top-1 routing reduces computational costs but also requires careful training to ensure balanced load distribution among the experts. To address this, we adopt load balancing ($\mathcal{L}_{LB}$) and router-z loss ($\mathcal{L}_{RZ}$) from recent works and report the results in row-3 of Table I. In addition to these, the model also requires a classification loss (i.e., $\mathcal{L}_C$) to optimize the model for defect identification. Since recent works also follow this route, we designate the MoE model using standard losses (i.e., $\mathcal{L}_C$+$\mathcal{L}_{LB}$+$\mathcal{L}_{RZ}$) as the baseline.

Importance of Our Proposed Loss Functions. We conduct an ablation study as shown in Table II to assess the importance of our proposed loss functions. We adopt two criteria to

perform a holistic ablation; namely criteria-1 and criteria-2. In criteria-1, we take a top-down approach by removing one loss at a time from combined standard and proposed losses, and measure the *increase* in the epochs required to train our model. Next, in criteria-2, we take a bottom-up approach, i.e., by adding one loss at a time to the standard loss functions and report the *reduction* in the epochs after adding the loss. Therefore, larger scores for both criterions signify a larger impact as both acknowledge the importance of the loss function under ablation. Our ablations show that $\mathcal{L}_E$ and $\mathcal{L}_G$ are most important among the four. Specifically, $\mathcal{L}_E$ has a score of 170 and 117 across criteria-1 and 2, respectively. Followed by $\mathcal{L}_G$ which has a score of 67 in criteria-1 and 219 in criteria-2. Finally, $\mathcal{L}_D$ and $\mathcal{L}_{ELB}$ have a score of 2 in criteria-1 whereas criteria-2 has 3 and 5, respectively. Although they don't show a large improvement, they still have a collaborative influence towards optimizing the model for faster convergence as reflected from results in Table I.

TABLE II
ABLATION STUDY ON OUR PROPOSED LOSS FUNCTIONS.

Loss Functions	Criteria-1 ↑	Criteria-2 ↑
L_E	**170**	117
L_G	67	**219**
L_D	2	3
L_{ELB}	2	5

Criteria-1: We ***remove*** one loss function at a time from proposed + standard losses and compute increase in epochs towards convergence to checkpoint-1.
Criteria-2: We ***add*** one loss function at a time from top-1 MoE to the standard loss functions and compute drop in epochs for convergence to checkpoint-1.
Interpretation: More (↑) the better for both criteria-1 and 2.

VI. CONCLUSION

Our work proposes tailored training objectives for sparse mixture-of-expert (MoE) models to facilitate faster convergence and improve expert utilization. We validated our approach by benchmarking against a strong baseline using current standard training objectives. We show that our method achieves 1.8% improvement in accuracy while using only 10% of training epochs (or training time) compared to the baseline. Furthermore, we support our results through a comprehensive ablation study and report the importance of our proposed approach. As evaluated on standards data, our proposed objectives for sparse MoE models mark a significant step in building scalable and efficient models for defect identification in non-destructive inspection.

VII. ACKNOWLEDGEMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, under contract $DE\text{-}AC02\text{-}06CH11357$.

REFERENCES

[1] Marija Bertovic and Iikka Virkkunen. "NDE 4.0: new paradigm for the NDE inspection personnel". In: *Handbook of Nondestructive Evaluation 4.0*. 2022, pp. 239–269.
[2] Fairouz Bettayeb, Tarek Rachedi, and Hamid Benbartaoui. "An improved automated ultrasonic NDE system by wavelet and neuron networks". In: *Ultrasonics* 42.1-9 (2004), pp. 853–858.
[3] Jian Chen, Zhenyang Yu, and Haoran Jin. "Nondestructive testing and evaluation techniques of defects in fiber-reinforced polymer composites: A review". In: *Frontiers in Materials* 9 (2022), p. 986645.
[4] William Fedus, Jeff Dean, and Barret Zoph. "A review of sparse expert models in deep learning". In: *arXiv preprint arXiv:2209.01667* (2022).
[5] William Fedus, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39.
[6] Sahar Hassani and Ulrike Dackermann. "A Systematic Review of Advanced Sensor Technologies for Non-Destructive Testing and Structural Health Monitoring". In: *Sensors* 23.4 (2023), p. 2204.
[7] Yash Jain et al. "DAMEX: Dataset-aware Mixture-of-Experts for visual understanding of mixture-of-datasets". In: *Advances in Neural Information Processing Systems* 36 (2024).
[8] AM Kokurov and DE Subbotin. "Ultrasonic detection of manufacturing defects in multilayer composite structures". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 1023. 1. 2021, p. 012013.
[9] Zachary Kral, Walter Horn, and James Steck. "Neural network approach of active ultrasonic signals for structural health monitoring analysis". In: *Health Monitoring of Structural and Biological Systems 2009*. Vol. 7295. SPIE. 2009, pp. 69–79.
[10] Marta Lagomarsino et al. "An online framework for cognitive load assessment in industrial tasks". In: *Robotics and Computer-Integrated Manufacturing* 78 (2022), p. 102380.
[11] Nicolas Leberruyer et al. "Toward Zero Defect Manufacturing with the support of Artificial Intelligence—Insights from an industrial application". In: *Computers in Industry* 147 (2023), p. 103877.
[12] Bo Li et al. "Sparse mixture-of-experts are domain generalizable learners". In: *arXiv preprint arXiv:2206.04046* (2022).
[13] Shaun McKnight et al. "A comparison of methods for generating synthetic training data for domain adaption of deep learning models in ultrasonic non-destructive evaluation". In: *NDT & E International* 141 (2024), p. 102978.
[14] Min Meng et al. "Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks". In: *Neurocomputing* 257 (2017), pp. 128–135.
[15] Navya Prakash et al. "Learning defects from aircraft NDT data". In: *NDT & E International* 138 (2023), p. 102885.
[16] Werickson FC Rocha et al. "Machine learning protocol from ultrasound data for monitoring, predicting, and supporting the analysis of dam slopes". In: *21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2022, pp. 1619–1623.
[17] Noam Shazeer et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer". In: *arXiv preprint arXiv:1701.06538* (2017).
[18] Majid Shirazi, Georgii Safronov, and Amr Rizk. "Multi-Modal Machine Learning for Navigating Noisy Objectives of Automotive Manufacturing Quality Inspection". In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. 2023, pp. 1875–1882.
[19] Rahul Soni et al. "A critical review of recent advances in the aerospace materials". In: *Materials Today: Proceedings* (2023).
[20] Hossein Towsyfyan et al. "Successes and challenges in non-destructive testing of aircraft composite structures". In: *Chinese Journal of Aeronautics* 33.3 (2020), pp. 771–791.
[21] Jiaxing Ye, Shunya Ito, and Nobuyuki Toyama. "Computerized ultrasonic imaging inspection: From shallow to deep learning". In: *Sensors* 18.11 (2018), p. 3820.
[22] Austin Yunker et al. "Comparative Study on Deep Learning Methods for Defect Identification and Classification in Composite Aerostructure Material". In: *50th Annual Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 87202. 2023, V001T05A001.
[23] Barret Zoph et al. "St-moe: Designing stable and transferable sparse expert models". In: *arXiv preprint arXiv:2202.08906* (2022).