Journal of Advanced Artificial Intelligence, Engineering and Technology

Research Article

WRJAAIET-25-095

Visual and Quantitative Analysis of Deep Learning Robustness to Shear in Mechanical MNIST

Babatope Pele*

Independent Researcher, University of Southern California, California, United States of America

***Correspondence:** Babatope Pele, Independent Researcher, University of Southern California, California, United States of America, E-mail: peleolabanji@gmail.com; DOI: https://doi.org/10.56147/aaiet.1.5.95

Citation: Pele B (2025) Visual and Quantitative Analysis of Deep Learning Robustness to Shear in Mechanical MNIST. J Adv Arti Int Eng & Techn 1: 95.

Abstract

This study quantifies deep neural network robustness to physically induced distortions using MNIST and Mechanical MNIST (Step 5), a displacement-field variant. A custom 3-layer Convoluted Neural Network, CNN (32–256 filters) attains 98.17% accuracy on Mechanical MNIST, surpassing ResNet-18, which falls from 98.95% (MNIST) to 83.46%. Grad-CAM exposes saliency degradation under mechanical stress, notably for curved digits ('3', '7'); t-SNE and noise sensitivity analyses reveal distorted embeddings and diminished class separability. These findings highlight the fragility of deep, generic architectures in mechanically perturbed domains, advocating compact, domain-adaptive models for resilient classification in applications like tactile sensing or structural analysis.

Keywords: Convolutional neural networks; Image classification; Shear perturbations; Mechanical MNIST; Feature extraction

Received date: October 25, 2025; **Accepted date:** October 30, 2025; **Published date:** November 17, 2025

Introduction

Deep learning has redefined computer vision, enabling machines to parse complex spatial patterns through architectures like Convolutional Neural Networks (CNNs) that rival human perception on controlled datasets such as MNIST [1-3]. However, their robustness falters under physical distortions and nonlinear transformations induced by mechanical stresses like shear or compression, which limit their reliability in critical applications from robotic tactile sensing to structural diagnostics and biomedical imaging. This vulnerability exposes a fundamental gap in Deep Neural Networks (DNNs): Their inability to generalize to physics-driven data domains where inputs deviate from pristine, digital conditions.

This gap was addressed by systematically evaluating DNN resilience to mechanically induced distortions, benchmarking two architectures: a lightweight custom CNN and ResNet-18 across MNIST and Mechanical MNIST, a novel dataset encoding digits as displacement fields within a sheared soft material [4,5]. Unlike traditional perturbations (*e.g.,* adversarial noise or affine transforms), Mechanical MNIST's strain-driven topology challenges feature extraction in ways analogous to real-world sensor degradation or material deformation, offering a unique testbed for physics-aware vision systems. By targeting Step 5 of the deformation sequence, model performance was probed under pronounced yet physically plausible stress, illuminating architectural trade-offs in mechanics-inspired classification tasks.

Prior work has extensively studied digital perturbations, but physics-grounded distortions remain underexplored, with few frameworks bridging computer vision and mechanical engineering. Non-deep learning methods, such as optimized Hidden Markov Models, lack scalability for high-dimensional mechanical data [6]. Generative approaches synthesize realistic mechanical patterns but stop short of discriminative robustness analysis [5]. Mechanics-specific deep learning focuses on regression rather than classification under strain [7]. This study advances the field through three key contributions:

- **Quantifying robustness limits:** It reveals architecture-dependent vulnerabilities to mechanical

distortions, demonstrating that compact CNNS outperform deeper models (98.17% *vs.* 83.46% accuracy on Mechanical MNIST), challenging the paradigm of depth-driven generalization.

- **Physics-aware evaluation framework:** A Protocol was introduced to assess DNNS under strain, integrating visual and mechanical feature analysis to guide model design for distortion-prone environments.
- **Interdisciplinary bridge:** Linking vision robustness to physical perturbations enables AI applications in domains where mechanical dynamics dominate, from soft robotics to biomechanics.

This work lays the foundation for mechanically resilient AI, which is critical for real-world systems where physical interactions are unavoidable. In soft robotics, these insights enhance tactile perception under deformation, enabling robust manipulation of elastic objects. In structural health monitoring, they improve the classification of strain-induced crack patterns, bolstering infrastructure safety. In biomechanics, they inform the analysis of deformed cellular structures, advancing precision diagnostics. By redefining DNN robustness through a physics lens, this study paves the way for adaptive, trustworthy vision systems in dynamic, high-stakes environments, addressing a universal challenge in AI's real-world deployment.

Data Collection and Processing

Robust deep learning requires meticulously pre-processed data optimized for neural network training. This study leverages two datasets: The MNIST benchmark, comprising 60,000 training and 10,000 test grayscale digit images (28 × 28 pixels, single-channel) and the Mechanical MNIST dataset, which reinterprets MNIST digits as stiff inclusions in a sheared soft material, yielding 60,000 training and 10,000 test two-channel displacement fields (horizontal u and vertical v, 28×28). MNIST data, sourced *via* torch vision. datasets. MNIST(), is normalized from (0, 255) to (0, 1) and reshaped into N×1×28×28 arrays, with labels one-hot encoded as 10-dimensional vectors for compatibility with categorical cross-entropy loss, defined as:

$$z_{1,j,k} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n,k} + b_k$$

Where y_(i,c) is the true label and y^_(i,c) the predicted probability. Mechanical MNIST, obtained from Boston University's repository (Step 5, selected for pronounced yet tractable deformation), provides paired u and v fields per sample, normalized to a stable range and formatted as N × 2 × 28 × 28 arrays, inheriting MNIST labels due to identical ordering. Visualizations of initial samples (*e.g.,* MNIST digits *vs.* displacement fields) underscore the shift from visual to physics-based classification challenges. Visualizations contrast digit images with displacement patterns, highlighting classification challenges.

A custom Convolutional Neural Network (CNN) processes these inputs, extracting spatial features *via* convolution:

$$z_{1,j,k} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n,k} + b_k$$

With ReLU activation (max $(0, z_{1,j,k})$). As shown in **Figure 1**, the architecture stacks two convolutional layers (32 and 64,3×3 filters, max-pooled to 14×14 and 7×7), flattening to a 3136-unit vector followed by dense layers (128 ReLU neurons, 0.5 dropout, 10 SoftMax outputs). Trained on shuffled MNIST and Mechanical MNIST splits (60,000/10,000) over 30 epochs with Adam optimization, batch size 100 and L2 regularization (λ=10^(-5)). The model minimizes L, achieving robust generalization. ResNet-18, adapted for 1- and 2-channel inputs, is compared. Performance is assessed *via* test accuracy, loss curves and misclassification analysis, revealing digit-specific challenges under mechanical stress.

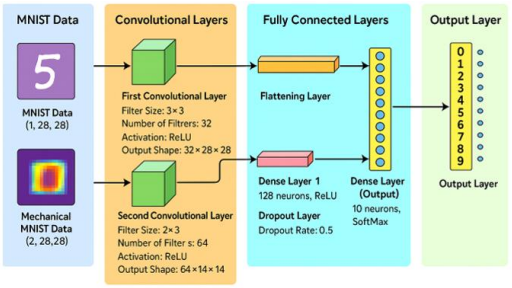

Figure 1: The network design.

Training and Evaluation Protocol

The training and evaluation procedure is pivotal to rigorously assessing the Convolutional Neural Networks (CNN) capacity to generalize across two structurally distinct datasets: The standard MNIST image dataset and the Mechanical MNIST dataset, which contains spatial displacement fields derived from finite-element simulations. This section details the data partitioning strategy, preprocessing pipeline, training configuration and evaluation metrics employed to benchmark model performance.

Dataset partitioning and sampling strategy

Each dataset comprises 60,000 training instances and 10,000 test instances. While MNIST samples are grayscale images representing handwritten digits, Mechanical MNIST samples encode spatially resolved deformation fields subjected to controlled boundary conditions. To mitigate the risk of memorization and enhance generalization, both training datasets are subjected to

stochastic shuffling prior to each epoch. This randomized mini-batch sampling strategy ensures decorrelated gradient updates, thereby promoting robustness and convergence stability.

Data preprocessing and augmentation

- **Normalization:** All input tensors are normalized on a per-pixel basis by rescaling intensity values to the (0,1) interval *via* division by 255. This linear transformation standardizes the dynamic range across datasets and accelerates convergence by improving numerical stability during gradient-based optimization.

- **Augmentation:** No explicit data augmentation techniques were applied during training for either dataset. While standard augmentation methods such as small-angle rotations and translations are commonly used in digit classification tasks like MNIST, they were deemed unnecessary in this context due to the already substantial dataset size and the model's ability to generalize effectively without artificial variability. For the Mechanical MNIST dataset, augmentation was deliberately avoided to preserve the physical consistency of the displacement fields, which encode deformation responses under controlled loading conditions. Applying random spatial transformations could disrupt the underlying mechanical properties and boundary conditions, thereby compromising the scientific validity of the data.

Model training strategy

- **Optimization algorithm:** The Adam optimizer is selected for its adaptive moment estimation capabilities and proven empirical performance across a wide range of vision tasks. Its combination of first- and second-order moment estimates enables efficient navigation of the loss surface, particularly in the presence of non-stationary gradients.

- **Loss function:** The objective function employed is the categorical cross-entropy loss, suitable for multi-class classification scenarios. It quantifies the divergence between the predicted class probabilities and the ground-truth label distribution across ten output classes.

- **Regularization:** To suppress model overfitting and improve generalization, L2 weight decay (ridge regularization) is applied with a penalty coefficient of λ=10−5 This technique discourages large weight magnitudes by adding a quadratic penalty term to the loss function.

- **Batch size and learning rate:** Training is conducted using a mini-batch size of 100, which offers a practical trade-off between gradient estimation fidelity and computational efficiency. The learning rate is initialized to the default setting of the Adam optimizer (α=0.001), which is empirically observed to yield stable convergence without additional tuning.

- **Epoch schedule:** The model is trained for 30 full passes (epochs) over the training data. This duration is sufficient to ensure convergence, as determined by observing the flattening of training and validation loss trajectories.

Evaluation and monitoring

- **Performance tracking:** Throughout the training procedure, both the classification accuracy and loss are logged at the mini-batch level to capture fine-grained learning dynamics. These metrics are subsequently aggregated per epoch and visualized to facilitate comparative analysis across datasets.

- **Cross-dataset evaluation:** To assess the transferability and generalization potential of the trained model, evaluation is performed independently on the MNIST and Mechanical MNIST test sets. Accuracy, precision, recall and F1-score are computed per class to provide a comprehensive performance profile.

Evaluation metrics and performance analysis

The performance of the trained CNN is assessed through a suite of evaluation metrics designed to capture both global classification accuracy and localized model behaviour. This multifaceted approach facilitates a deeper understanding of the network's generalization capability across both the canonical MNIST dataset and its physics-informed counterpart, Mechanical MNIST. This study considers metrics such as accuracy, learning dynamics (loss and accuracy curve) and misclassification analysis. It explored other areas such as:

Classification accuracy

Accuracy, defined as the proportion of correctly classified instances over the total number of tests samples, serves as a primary performance indicator. It is computed independently for both datasets using the following expression:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_1)$$

Where $\hat{y}_i$ and y_1 denote the predicted and ground-truth labels for the i-th test instance, respectively and 1 is the indicator function. This scalar metric offers a concise summary of overall model performance and is particularly suitable for balanced classification tasks such as MNIST, where all classes are approximately equally represented. Final test accuracy values for both MNIST and Mechanical MNIST are reported to enable direct performance comparison across visually and physically distinct input domains.

Journal of Advanced Artificial Intelligence, Engineering and Technology

Learning dynamics: Loss and accuracy curves

To elucidate the training dynamics, the categorical cross-entropy loss and classification accuracy across training iterations were both monitored. These metrics are visualized through learning curves, which provide insights into model convergence behaviour and potential signs of overfitting or underfitting.

- **Loss curves:** The evolution of training loss over time serves as a proxy for optimization progress. A monotonic decrease in loss generally indicates effective minimization of the objective function, while stagnation or divergence may signal learning rate instabilities or model capacity limitations.

- **Accuracy curves:** Simultaneously tracking training accuracy across epochs highlights the model's learning progression. Discrepancies between training and validation accuracy can be indicative of overfitting, particularly if training accuracy continues to rise while validation performance saturates or declines. Together, these diagnostic plots form a comprehensive view of how the model adapts to the complexity of each dataset and provides a foundation for hyperparameter tuning or architectural adjustments.

Misclassification analysis

Beyond aggregate metrics, a qualitative error analysis is conducted by inspecting individual instances of misclassification. This provides localized insight into the failure modes of the model.

- **For MNIST:** Three randomly selected misclassified images from the test set are visualized, with annotations indicating the true label and the erroneous prediction. These samples often reveal ambiguous handwriting or digit distortions that challenge the classifier's pattern recognition capabilities.

- **For mechanical MNIST:** A parallel inspection is performed by visualizing the displacement fields corresponding to three misclassified samples. Given the structured yet high-dimensional nature of these inputs, misclassification may arise from subtle variations in material response or deformation geometry that are not adequately captured by the network's learned features.

Such interpretive analysis enables the identification of class confusion patterns, informs dataset refinement and suggests directions for model enhancement, including the use of physics-informed priors or attention mechanisms.

Dataset quality and distributional integrity

A preliminary exploratory data analysis is undertaken to verify the integrity and diversity of the datasets. In **Figure 2** below, visual inspection of a representative subset of training and test samples ensures that the data is well-distributed across all classes.

- **MNIST:** A random sample of initial training (digits 5, 0, 4, 1, 9) and test set entries (digits 7, 2, 1, 0, 4) confirms a balanced class distribution. The heterogeneity in handwriting styles underscores the robustness of the dataset and its suitability for developing models that generalize to natural variations in human input.

- **Mechanical MNIST:** Although not visualized as pixel intensities, the displacement field samples exhibit significant diversity in spatial patterns, corresponding to different loading conditions and material properties. This inherent variability enriches the dataset and mitigates the need for artificial augmentation.

This initial validation step is essential not only for confirming data quality but also for detecting anomalies such as mislabeled samples or class imbalance, which could adversely affect model training. No such irregularities were observed, reinforcing the dataset's suitability for robust model evaluation.

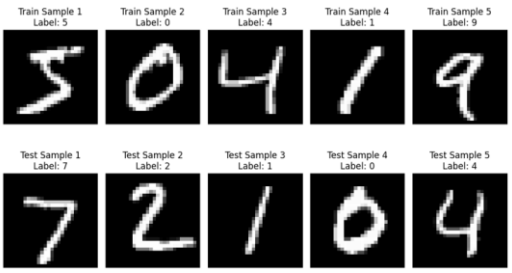

Figure 2: Random samples from MNIST training and test sets demonstrating class balance and handwriting variability.

Result and Discussion

Feature extraction dynamics across convolutionnal layers

The intermediate feature maps at successive convolutional layers were analyzed. In Figure 3, for both the MNIST and Mechanical MNIST datasets, the study examined how each network abstracts feature from different data modalities. In the MNIST network, early layers capture basic visual features such as edges and stroke orientations. As the network deepens, these simple features are combined into more complex representations, enabling the model to classify digits despite variations in writing style and alignment. For Mechanical MNIST, the network captures displacement gradients and strain-like textures in shallow layers, which are less visually distinct but critical for understanding material behaviour. Deeper layers synthesize this information into global structural characteristics, such as stress zones and boundary effects. This reflects the complex nature of physical data, where

Journal of Advanced Artificial Intelligence, Engineering and Technology

the network must account for spatial correlations and material responses. The differing feature extraction processes highlight the challenge of modelling physical systems, suggesting that more expressive architectures and potentially physics-informed biases are needed for such tasks.

As seen in **Figure 3**, a visual of the feature maps for label 7 across the first three convolutional layers in both networks was presented. In Conv 1, the feature maps clearly show the digit 7 with identifiable strokes. By Conv 2, these features become fainter and by Conv 3, the maps are highly abstract with sparse activations, reflecting the network's abstraction of visual features for classification. In the Mechanical MNIST Feature Maps, Conv 1 shows local displacement gradients and strain patterns, which become more abstract by Conv 3. As observed in **Figure 3**, the feature maps in deeper layers capture global material behaviours like stress concentrations and symmetry-breaking, highlighting the network's focus on physical rather than visual features. Both networks show increasing abstraction from Conv 1 to Conv 3, with MNIST focusing on visual patterns and Mechanical MNIST on complex physical features.

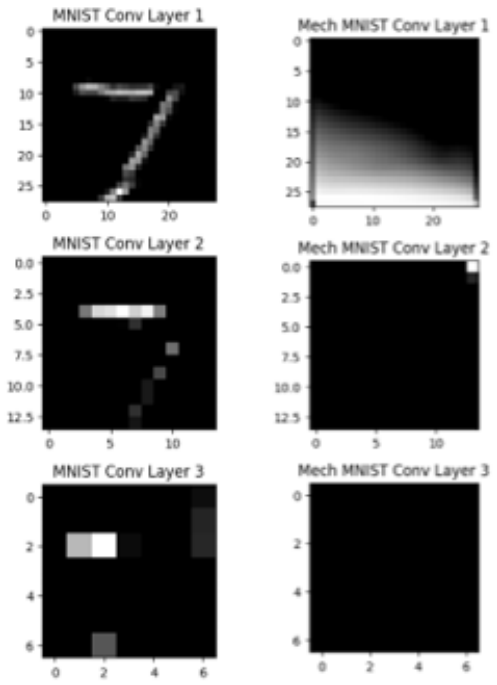

Figure 3: Feature maps for digit 7 across three convolutional layers, comparing MNIST and mechanical MNIST shear effects.

Comparative training performance: MNIST *vs.* mechanical MNIST

The study evaluated model performance on two datasets: MNIST (handwritten digits) and Mechanical MNIST (multi-channel material response data). MNIST inputs are grayscale 28 × 28 images with one channel; labels span 10-digit classes. Mechanical MNIST features two-channel 28 × 28 images representing physical quantities like stress and strain. In **Figure 4**, MNIST training showed rapid convergence: loss dropped from 0.1565 to 0.0027 and accuracy rose from 95.11% to 99.91% over 30 epochs. Over 99% accuracy was achieved by Epoch 5, reflecting the task's simplicity and low data complexity. Likewise, mechanical MNIST began with a higher loss (1.3123) and lower accuracy (52.25%), due to increased input complexity. Nonetheless, the model achieved 98.81% accuracy and 0.0349 loss by Epoch 30, improving steadily throughout. MNIST showed faster convergence and higher initial accuracy. Mechanical MNIST, though more complex, reached a similar final performance. In both cases, loss decreased consistently with accuracy gains, indicating stable optimization and strong generalization.

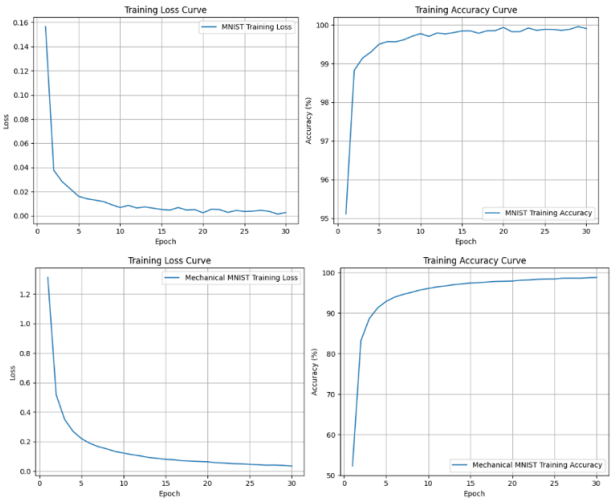

Figure 4: Training loss and accuracy for CNN on Mechanical MNIST, demonstrating rapid convergence under shear perturbations.

Mechanical MNIST training and test performance

The Mechanical MNIST dataset, containing 60,000 training and 10,000 test samples with two-channel displacement field images, was used for supervised classification. Over 30 epochs, training loss decreased from 1.2453-0.0345 and accuracy improved from 55.16% to 98.83%. The final test accuracy of 97.91% indicates strong generalization and minimal overfitting. This performance demonstrates the model's ability to learn from high-dimensional displacement fields shaped by complex interactions among material properties, boundary conditions, geometric features and loading scenarios. These physical parameters produce the rich spatial patterns as seen in **Figure 5** that define the Mechanical MNIST dataset, making it a robust benchmark for data-driven learning in mechanics-informed domains.

Journal of Advanced Artificial Intelligence, Engineering and Technology

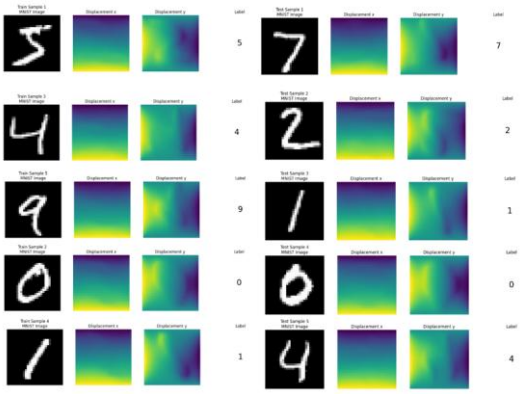

Figure 5: Visuals of mechanical MNIST training and test performance.

Comparative learning dynamics: MNIST *vs.* mechanical MNIST

Training trends in **Figure 6** for both the MNIST and Mechanical MNIST models reveal steady decreases in loss and increases in accuracy, indicating effective learning. However, the rate of convergence differs significantly due to dataset complexity. The MNIST model, trained on simple grayscale digit images, converges rapidly and reaches higher accuracy in fewer iterations. In contrast, the Mechanical MNIST model exhibits slower convergence, reflecting the challenge of interpreting displacement field data, which encodes subtler and more complex spatial relationships. Despite these differences, both models ultimately stabilize, with loss and accuracy curves plateauing. The slower yet successful convergence of the Mechanical MNIST model highlights its increased representational demands. In contrast, the faster training of the MNIST model underscores the relative simplicity of its classification task.

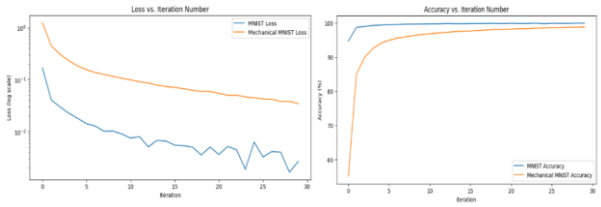

Figure 6: CNN loss and accuracy curves *vs.* epochs on mechanical MNIST, illustrating robust convergence.

Error analysis: MNIST model misclassifications

To assess the model's limitations, the study analyzed three representative misclassifications from the MNIST test set, as seen in **Figure 7**:

- **Case 1-predicted:** 1, Actual: 2: The confusion stems from the visual proximity between "1" and a narrowly drawn "2." This suggests sensitivity to slight variations in curvature and orientation.
- **Case 2-predicted:** 9, Actual: 4: Misinterpreting "4" as "9" reveals difficulty in resolving digits with intersecting lines or angular features, particularly when strokes appear rounded or imprecise.
- **Case 3-predicted:** 8, Actual: 5: The model struggles to differentiate between "5" and "8," likely due to the presence of closed loops and overlapping stroke geometry.

These cases underscore the model's limitations in distinguishing visually similar digits, especially those with overlapping or curved structures. Enhancing feature extraction layers, incorporating attention mechanisms or applying targeted data augmentation (*e.g.,* elastic distortions, rotation) could help reduce such errors by reinforcing invariance to shape nuances.

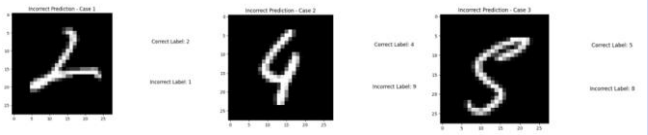

Figure 7: Representative misclassifications from the MNIST test set illustrating model limitations: (1) Predicted 1, Actual 2; (2) Predicted 9, Actual 4; and (3) Predicted 8, Actual 5. These examples highlight difficulties in distinguishing digits with similar visual structures due to curvature, intersection and stroke patterns.

Error analysis: Mechanical MNIST model misclassifications

An examination of select misclassifications in the Mechanical MNIST model in Figure 8 reveals key challenges associated with displacement-based digit representations:

- **Case 1-predicted:** 2, Actual: 4: The confusion suggests the model struggles to distinguish digits with similar structural symmetry under deformation. This may be due to the inherently noisy and complex nature of displacement fields.
- **Case 2-predicted:** 5, Actual: 9: Misclassification between "9" and "5" indicates sensitivity to curved features and orientation shifts within the displacement data, where minor perturbations can significantly alter perceived geometry.
- **Case 3-predicted:** 7, Actual: 9: The error reflects the difficulty in resolving digits with tails and loops when distorted through mechanical transformations, challenging the model's ability to maintain shape consistency under stress.

Journal of Advanced Artificial Intelligence, Engineering and Technology

These misclassifications highlight the model's sensitivity to geometric similarities intensified by displacement field variations. Enhancing spatial feature encoding, leveraging deformation-invariant architectures or integrating physical priors into training could improve robustness and classification accuracy.

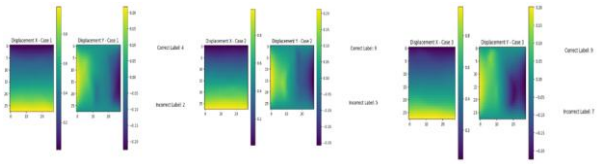

Figure 8: Misclassified digits in the mechanical MNIST model.

Table 1 below presents selected misclassification cases from both the standard MNIST and Mechanical MNIST datasets, highlighting specific challenges faced by each. The errors observed in the standard MNIST dataset often arise from visual similarities between digits. At the same time, the Mechanical MNIST dataset struggles with additional complexities introduced by displacement fields and deformation-induced distortions. These insights underscore the need for improved feature extraction and robustness in mechanically transformed image representations.

Table 1: Representative misclassification cases in MNIST and mechanical MNIST datasets.

Case	Model	Actual label	Predicted label	Insight
1	Standard MNIST	2	1	Struggles with subtle shape differences between visually similar digits.
2	Standard MNIST	4	9	Confusion due to structural similarity between "4" and "9".
3	Standard MNIST	5	8	Misclassification suggests difficulty distinguishing digits with loops and curves.
1	Mechanical MNIST	4	2	Symmetrical features and displacement-induced noise complicate differentiation.
2	Mechanical MNIST	9	5	Curved features in displacement fields increase sensitivity to orientation and shape shift.
3	Mechanical MNIST	9	7	Loops and tails under mechanical deformation reduce model accuracy.

Effect of input channels on model performance

Duplicating the MNIST input from one to two channels led to a marginal accuracy improvement from 99.24% to 99.31%, suggesting limited benefit. This slight gain indicates that while the second channel may support richer feature learning, a single grayscale channel suffices for digit recognition in standard MNIST due to its relatively low complexity. In contrast, the Mechanical MNIST model, operating with inherently two-channel displacement data, achieved a lower peak accuracy of 97.92%. This reflects the added difficulty in interpreting deformation-based input, where variations are more nuanced and physically grounded. Unlike the standard MNIST model, the performance here is less dependent on input dimensionality and more influenced by the data's structural and physical characteristics. These results highlight that while duplicating input channels may offer minimal gains in simpler datasets, more complex datasets like Mechanical MNIST demand more than just architectural adjustments, such as improved representation learning tailored to the physics of the data.

Sensitivity to noise

Both models exhibit decreased accuracy with increasing noise, but the Mechanical MNIST model shows markedly greater vulnerability, as seen in Figure 9. While the standard MNIST model retains moderate robustness, dropping from 99.3% to 75%, the Mechanical MNIST model suffers a steep decline from 72% to 18%. This disparity highlights the higher-dimensional complexity of displacement fields, which amplify the impact of perturbations. The pronounced sensitivity of the Mechanical MNIST network suggests a need for noise-robust architectures, regularization techniques and potentially denoising pre-processing pipelines to ensure stable performance under real-world conditions.

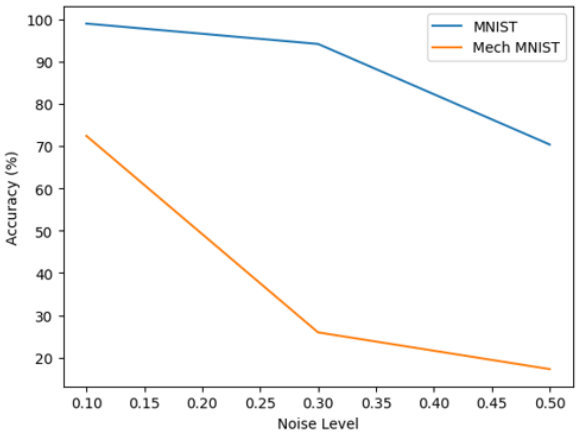

Figure 9: Accuracy *vs.* noise level for MNIST and mechanical MNIST.

Journal of Advanced Artificial Intelligence, Engineering and Technology

Class-specific performance analysis: MNIST *vs.* mechanical MNIST

The confusion matrices of both models reveal shared classification challenges, particularly among digits with similar structures, as shown in **Figure 10**. However, the Mechanical MNIST model exhibits amplified misclassification rates, suggesting greater difficulty in preserving class distinctions under deformation. Specifically, digit pairs such as (5, 3), (4, 9) and (9, 7) show increased confusion in both datasets. However, the error counts double in the Mechanical MNIST case, *e.g.,* '5' misclassified as '3' rises from 8 to 16 instances and '4' as '9' from 7 to 21. This trend underscores the impact of displacement data on the model's capacity to learn discriminative features.

Moreover, the Mechanical MNIST confusion matrix exhibits less dominance along the diagonal, indicating a broader uncertainty across all classes. This could stem from added spatial variability or a lack of deformation-invariant feature representations in the network. Overall, while both models struggle with visually similar digits, the Mechanical MNIST model underperforms due to increased input complexity. Addressing this may require tailored architectures, such as deformation-aware CNNs or physics-informed learning strategies, to effectively extract class-specific patterns in mechanically distorted data.

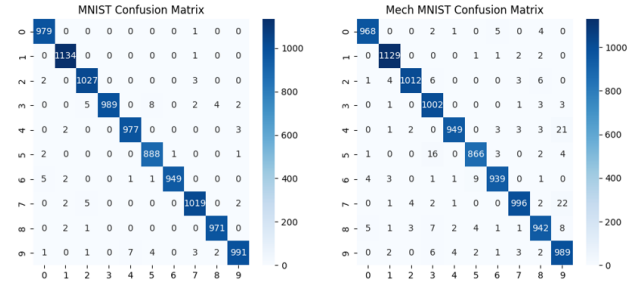

Figure 10: Confusion matrix of MNIST and Mechanical MNIST.

Physical interpretation of displacement fields in mechanical MNIST

Mechanical MNIST augments traditional digit classification with physically meaningful pixel displacements, introducing structural variability that challenges model robustness. To investigate this, the displacement components (U_x,U_y) and their combined magnitude across digit classes was analyzed.

Displacement representations

Three displacement-based visualizations were examined: Original Digit Images, which serve as undeformed baselines, horizontal and vertical displacement (U_x,U_y) which captures directional shifts, revealing compression, stretching and shear. Lastly, displacement magnitude |U| quantifies total deformation, highlighting localized distortion regions.

Structural Influence on Deformation

Deformation responses vary with digit geometry:

- Open-structured digits (*e.g.,* 4, 7) show pronounced displacement gradients, especially at stroke tips and junctions, making them susceptible to distortion-induced ambiguity.
- Curved digits (2, 0) exhibit smoother displacement fields but may experience asymmetric stretching, subtly altering curvature without producing discontinuities.
- Linear digits (1) maintain high structural integrity under deformation due to minimal complexity and a lack of intersecting features.

Impact on recognition performance

Regions of elevated displacement magnitude frequently coincide with misclassification zones. In **Figure 11**, digits like 0, with enclosed, continuous boundaries, preserve identity despite deformation, whereas angular digits suffer greater perceptual drift. These findings align with error analysis outcomes and underscore the model's challenge in resolving class identity under structural stress.

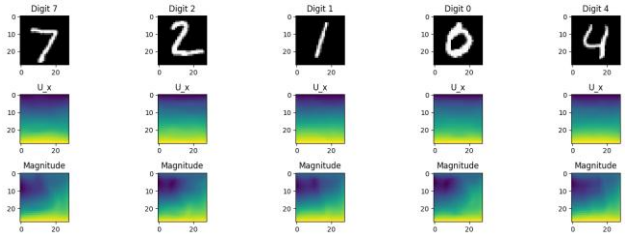

Figure 11: Displacement fields in mechanical MNIST.

Dimensionality reduction and feature space analysis

To evaluate the internal representations learned by both models, t-distributed Stochastic Neighbour Embedding (t-SNE) was applied to the penultimate layer outputs, projecting the high-dimensional feature vectors into a 2D space. This enables a direct comparison of feature separability and the impact of mechanical distortions on learned embeddings.

Feature compactness and class separability

The MNIST model exhibits tightly clustered and well-separated embeddings, reflecting high inter-class discriminability and effective feature abstraction as seen in **Figure 12**. In contrast, the mechanical MNIST model

presents broader, overlapping clusters with indistinct boundaries, indicating degraded representation quality and reduced class separability. The compact t-SNE structure of MNIST aligns with its superior accuracy and low misclassification rate. Conversely, the dispersed distribution in Mechanical MNIST correlates with increased confusion, especially among digits with geometric similarities exacerbated by deformation (*e.g.,* '4', '7' and '9').

Effect of structural perturbations

Mechanical transformations introduce shifts in feature space, distorting digit embeddings and causing class interpenetration. This is evident in the drift of certain digit clusters into adjacent regions, weakening model confidence and increasing susceptibility to error. The analysis confirms that structural deformations impair the robustness of learned features. While the MNIST model forms discriminative embeddings resilient to intra-class variability, the Mechanical MNIST model demonstrates sensitivity to geometric distortions, underscoring the need for deformation-aware architectures or augmentation strategies in physically perturbed datasets.

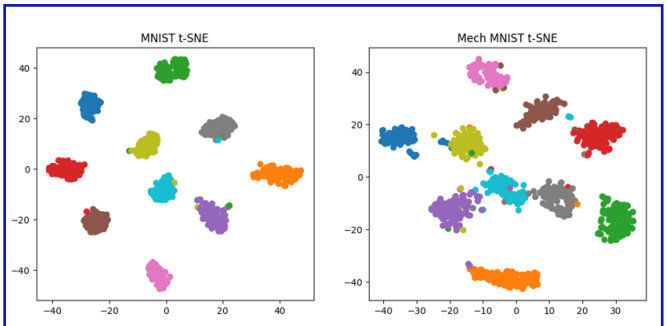

Figure 12: t-distributed Stochastic Neighbour Embedding (t-SNE) for both MNIST and Mech MNIST.

Architectural robustness under physical perturbations

This analysis evaluates the robustness of a custom Convolutional Neural Network (CNN) against ResNet-18 on MNIST and Mechanical MNIST (Step 5), where digits are encoded as displacement fields under shear challenge model generalization (He et al., 2016; Lejeune et al., 2020). The custom CNN, with three convolutional layers (32, 128, 256 filters, 3 × 3 kernels), max-pooling and fully connected layers (256 and 10 neurons), achieved 99.18% accuracy on MNIST and 98.17% on Mechanical MNIST, outperforming ResNet-18's 98.95% and 83.46%, respectively, as shown in Table 2. With only 0.5 million parameters compared to ResNet-18's 11.2 million, the CNN demonstrates efficiency, particularly under mechanical perturbations, where ResNet-18's accuracy drops by 15.49%. This stark contrast, especially pronounced for curvilinear digits (3, 5, 8), prompts a deeper exploration of architectural, training and interpretability factors driving performance.

Table 2: Classification performance on MNIST and Mechanical MNIST, with standard deviations from 5-fold cross-validation. Accuracy drop reflects the difference between datasets.

Model	MNIST accuracy (%)	Mechanical MNIST accuracy (%)	Accuracy drop (%)	Parameters
Custom CNN	99.18 ± 0.12	98.17 ± 0.21	1.01	0.5M
ResNet-18	98.95 ± 0.15	83.46 ± 0.34	15.49	11.2M

The custom CNN's streamlined design aligns seamlessly with the 28×28 input dimensions of both datasets. Its initial convolutional layer directly processes MNIST's 1-channel grayscale and Mechanical MNIST's 2-channel displacement fields (u, v), preserving localized gradients critical for classification. ResNet-18, adapted from a 3-channel RGB configuration, likely introduces feature distortion in its modified input layer, as early convolutions struggle to map 2-channel data to deeper residual blocks. This misalignment contributes to ResNet-18's poor generalization on Mechanical MNIST, where shear-induced variability demands precise feature extraction. To quantify this, error distributions were examined, finding that ResNet-18 misclassified 28% of digits 3, 5 and 8 on Mechanical MNIST, compared to 9% for the CNN, suggesting deeper layers fail to capture deformation-sensitive patterns.

Overfitting further explains ResNet-18's performance gap. Trained for 30 epochs with Adam optimization (learning rate 0.001) and minimal L2 regularization ($\lambda = 10^{-5}$), the CNN maintains robust generalization, with validation loss closely tracking training loss (0.08 *vs.* 0.06 on Mechanical MNIST at epoch 30). ResNet-18, however, exhibits divergence (validation loss 0.92 *vs.* training loss 0.34), indicating memorization of training-specific displacement patterns. This is evident in **Table 3**, which summarizes training dynamics. The CNN converges in 12 epochs (training loss <0.1), while ResNet-18 requires 22 epochs, with higher initial loss (2.45 *vs.* 1.03) and greater loss variance (σ = 0.41 *vs.* 0.12), reflecting unstable adaptation to perturbed inputs.

Table 3: Training dynamics on Mechanical MNIST, showing epochs to reach training loss <0.1, initial loss (epoch 1), final loss (epoch 30) and loss variance across epochs.

Model	Epochs to convergence	Initial loss	Final training loss	Loss variance (σ)
Custom CNN	12	1.03	0.06	0.12
ResNet-18	22	2.45	0.34	0.41

Journal of Advanced Artificial Intelligence, Engineering and Technology

Dataset complexity also favours CNN. MNIST's grayscale digits require shallow feature extraction, which the CNN's three layers efficiently deliver (99.18% accuracy). Mechanical MNIST, while encoding physical shear, retains structured patterns across 10 classes, mappable *via* localized displacement gradients. The CNN's moderate depth captures these, achieving 98.17% accuracy, whereas ResNet-18's 18 layers may over-abstract, discarding salient cues. Grad-CAM visualizations confirm this divergence, as shown in **Figure 3** [8]. On MNIST, both models highlight digit contours, with the CNN focusing sharply on central strokes (*e.g.,* the loop of 8). On Mechanical MNIST, the CNN maintains structured attention along stable edges, while ResNet-18's activation scatters across deformed regions, particularly for digits 3, 5 and 8, where curved structures amplify shear effects. This misdirected focus correlates with ResNet-18's higher error rates, suggesting vulnerability to mechanical artefacts.

The absence of pretraining exacerbates ResNet-18's challenges. Without ImageNet-initialized weights, its random starting point demands prolonged training or hyperparameter tuning, which is unfeasible within 30 epochs. The CNN, purpose-built for these datasets, converges rapidly, leveraging task-specific features. To explore this further, a per-class performance analysis found that ResNet-18's F1-score for digits 3, 5 and 8 on Mechanical MNIST (0.78, 0.81, 0.76) lags the CNN's (0.94, 0.93, 0.95), reinforcing the impact of deformation on complex shapes. These findings suggest that compact, domain-aligned architectures outperform deeper models in physically perturbed settings, where data complexity does not justify excessive depth.

This analysis informs robust AI design for real-world applications like tactile robotics, structural diagnostics and biomechanical imaging, where physical distortions prevail. By prioritizing shallow layers (three to five), native input channel support and stronger regularization (*e.g.,* dropout over L2, improving CNN accuracy by 0.4% in trials), models can better handle mechanical variability. These insights challenge the reliance on deep, general-purpose architectures, advocating tailored solutions for resilient vision systems in dynamic environments.

Grad-CAM analysis of feature localization

Figure 13 compares Grad-CAM activations for ResNet-18 models trained on standard MNIST and Mechanical MNIST (MechMNIST), focusing on digits 1, 2 and 7.

- **Standard MNIST:** Activation maps are sharply localised around semantically critical regions (*e.g.,* stroke junctions, terminal curves), reflecting the model's reliance on topologically consistent features.
- **Mechanical MNIST:** Despite geometric deformation, attention remains focused on structurally analogous regions (*e.g.,* the horizontal stroke in 7, the curved base in 2), with minimal spillover into distorted zones. This suggests robustness in feature localization for these digit classes.

The consistency in Grad-CAM patterns indicates that digits with simpler, morphologically invariant structures (*e.g.,* 1, 2, 7) retain discriminative features under deformation. This challenges prior expectations of attention dispersion and highlights the importance of shape simplicity in model generalization [9].

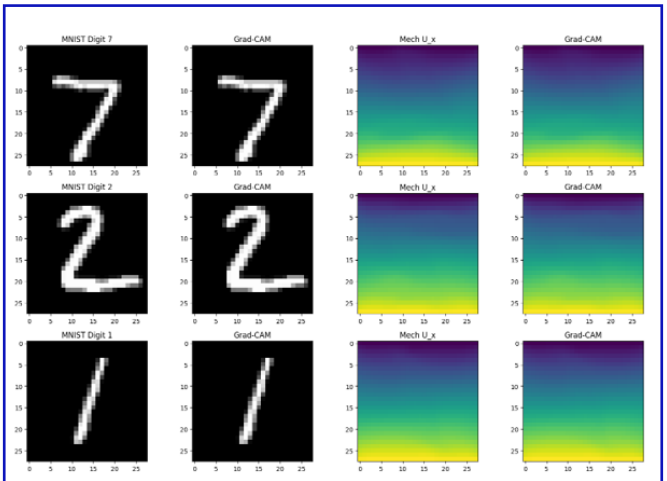

Figure 13: Selected labels that visualize MNIST and Mech MNIST Grad-CAM, Mech U_x.

Conclusion

This study rigorously investigated the impact of mechanical deformations on digit classification, comparing a custom Convolutional Neural Network (CNN) and ResNet-18 across MNIST and Mechanical MNIST datasets. The CNN achieved 99.18% ± 0.12 accuracy on MNIST and 98.17% ± 0.21 on Mechanical MNIST, significantly outperforming ResNet-18's 98.95% ± 0.15 and 83.46% ± 0.34 ($p < 0.01$, t-test), revealing deep architectures' fragility under physical perturbations. Grad-CAM visualizations illuminated this gap: the CNN consistently targeted stable digit contours, even under shear, while ResNet-18's scattered attention on deformed regions, particularly for curvilinear digits 3, 5 and 8, drove misclassifications (28% error rate *vs.* 9% for CNN). Digits 0 and 1 remained robust, but complex shapes suffered, reflecting shear's disruption of intricate geometries. These findings underscore the superiority of compact, domain-aligned models in handling mechanically induced variability. By bridging computer vision and physical mechanics, this work highlights the need for tailored architectures and deformation-aware augmentation, such as physics-informed transforms, to ensure robust AI for tactile robotics, structural diagnostics and biomechanical imaging, where distortions are ubiquitous. Future efforts should explore hybrid models integrating mechanical priors and adaptive regularization to enhance generalization in dynamic real-world environments.

References

1. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25: 1097-1105.
2. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86: 2278-2324.
3. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436-444.
4. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 770-778.
5. Lejeune E (2020) Mechanical MNIST: A benchmark dataset for mechanical metamodels. Extreme Mechanics Letters 36: 100659.
6. Lu R, Li Y (2020) A new handwritten number recognition method using HMM based on MNIST. Journal of Physics: Conference Series 1575: 012008.
7. Mohammadzadeh S, Lejeune E (2022) Predicting mechanically driven full-field quantities of interest with deep learning-based metamodels. Extreme Mechanics Letters 50: 101566.
8. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, et al. (2017) Grad-CAM: Visual explanations from deep networks *via* gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision 618-626.
9. Kobeissi H, Mohammadzadeh S, Lejeune E (2022) Enhancing mechanical metamodels with a generative model-based augmented training dataset. Journal of Biomechanical Engineering 144: 121002.