International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Steel Surface Defect Detection Using Deep Learning: A Comparative Study of MobileNetV3Large and VGG19+InceptionV3 Ensemble

Sanjeev Kumar

Research Scholar, Department of Computer Science and Engineering, A.N.A College of Engineering & Management, Bareilly

Dr. Vineet Agarwal

Professor, Department of Computer Science and Engineering, A.N.A College of Engineering & Management, Bareilly

ABSTRACT

Automated surface defect detection in steel manufacturing is a critical quality-control challenge that directly impacts production efficiency and product reliability. This paper presents a comprehensive deep learning-based approach for classifying six types of steel surface defects—crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches—using the NEU Surface Defect Database (1,440 images). We evaluate two architectures: (1) MobileNetV3Large with transfer learning, and (2) a dual-input ensemble of VGG19 and InceptionV3. An advanced preprocessing pipeline (CLAHE, Gaussian denoising, unsharp masking, normalization) and extensive data augmentation are applied. The ensemble achieves 99.07% accuracy, 99.10% precision, 99.07% recall, and 99.07% F1-score, outperforming MobileNetV3Large (89.81% accuracy) by ~9.3 percentage points. LIME explainability analysis validates the model's focus on semantically meaningful defect regions, making the framework suitable for industrial deployment.

Keywords:- Surface defect detection, deep learning, transfer learning, MobileNetV3Large, VGG19, InceptionV3, ensemble learning, CLAHE, LIME, NEU dataset.

I. INTRODUCTION

Steel surface quality is decisive for the structural integrity and safety of products in automotive, aerospace, and construction industries. Defects introduced during hot-rolling—crazing, inclusions, patches, pitted surfaces, rolled-in scales, and scratches—must be detected reliably and rapidly. Manual inspection is labor-intensive, subjective, and inconsistent at production-line speeds, motivating the adoption of automated vision-based inspection systems.

Convolutional neural networks (CNNs) have demonstrated exceptional capability in image classification, consistently outperforming handcrafted feature-based approaches. Pretrained architectures—MobileNetV3, VGG19, InceptionV3—trained on ImageNet provide powerful transferable features adaptable to industrial defect domains. Ensemble methods combining multiple backbones further improve robustness by leveraging complementary representations.

This paper contributes: (1) an advanced image preprocessing pipeline tailored for steel surface imagery; (2) a systematic evaluation of MobileNetV3Large vs. a VGG19+InceptionV3 dual-input ensemble; (3) comprehensive quantitative analysis across accuracy, precision, recall, F1-

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

score, ROC curves, PR curves, confusion matrices, and confidence distributions; and (4) LIME-based explainability for industrial trust.

II. RELATED WORK

Early defect detection relied on handcrafted features (HOG, LBP, Gabor) with SVM or Random Forest classifiers. While reasonable on controlled datasets, these methods struggled with illumination variability and morphological diversity. Shi et al. [1] showed CNN-based features significantly outperform handcrafted alternatives on steel strip defects. Song et al. [2] demonstrated ResNet architectures achieving >92% accuracy on the NEU dataset via residual connections.

Transfer learning has become dominant for defect detection where labeled data is scarce. Cha et al. [4] applied VGG-based ImageNet features to crack detection. Lightweight models (MobileNetV2, EfficientNet) enable edge deployment [5]. Ensemble methods—combining multiple pretrained CNNs—have shown further gains. Tabernik et al. [6] proposed a segmentation-based ensemble for unsupervised defect localization. Our work extends this body of literature with a dual-input ensemble specifically benchmarked on NEU with full explainability analysis.

III. DATASET DESCRIPTION

The NEU Surface Defect Database [7] contains 1,440 grayscale hot-rolled steel strip images, uniformly distributed across six defect classes (240 images each): crazing (Cr), inclusion (In), patches (Pa), pitted surface (PS), rolled-in scale (RS), and scratches (Sc). The balanced distribution prevents class-imbalance bias. Fig. 1 shows the dataset class distribution and sample images across all defect types.

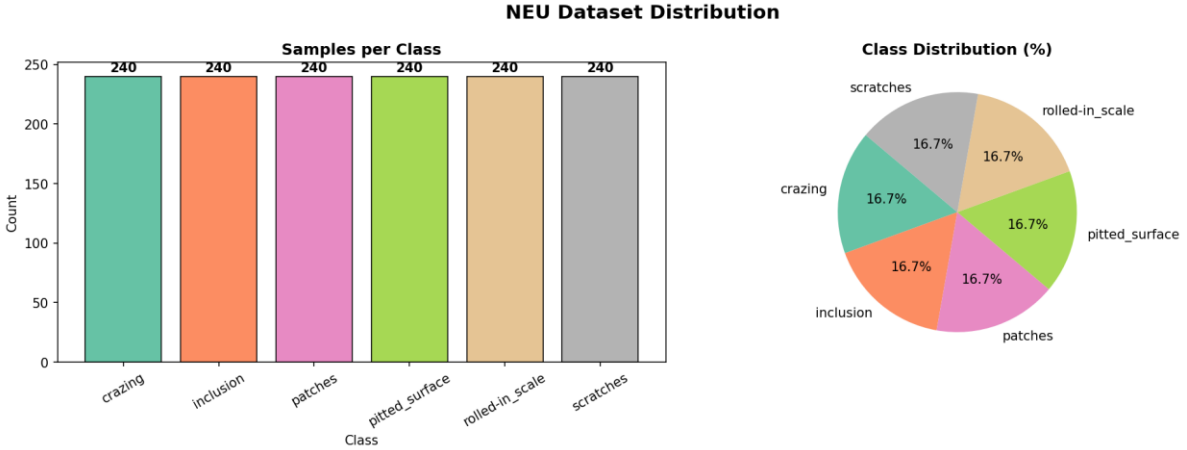

Fig. 1. NEU Surface Defect Dataset Distribution — Bar chart (left) and pie chart (right) showing equal distribution of 240 images per class, totaling 1,440 images.

Table I: NEU Surface Defect Dataset Distribution

Defect Class	Images	% of Total	Abbrev.
Crazing	240	16.67%	Cr

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Inclusion	240	16.67%	In
Patches	240	16.67%	Pa
Pitted Surface	240	16.67%	PS
Rolled-in Scale	240	16.67%	RS
Scratches	240	16.67%	Sc
Total	**1,440**	**100.00%**	—

IV. METHODOLOGY

A. Image Preprocessing Pipeline

Raw steel images exhibit low defect-background contrast, rolling process noise, and subtle textures. Our multi-stage pipeline addresses these challenges systematically. Fig. 3 illustrates the complete preprocessing pipeline with before/after comparisons.

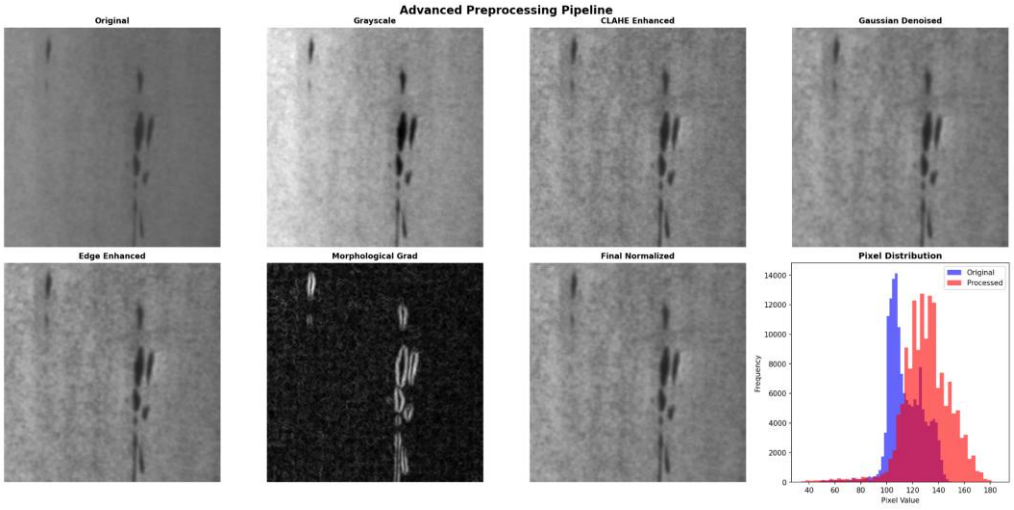

Fig. 3. Advanced Preprocessing Pipeline

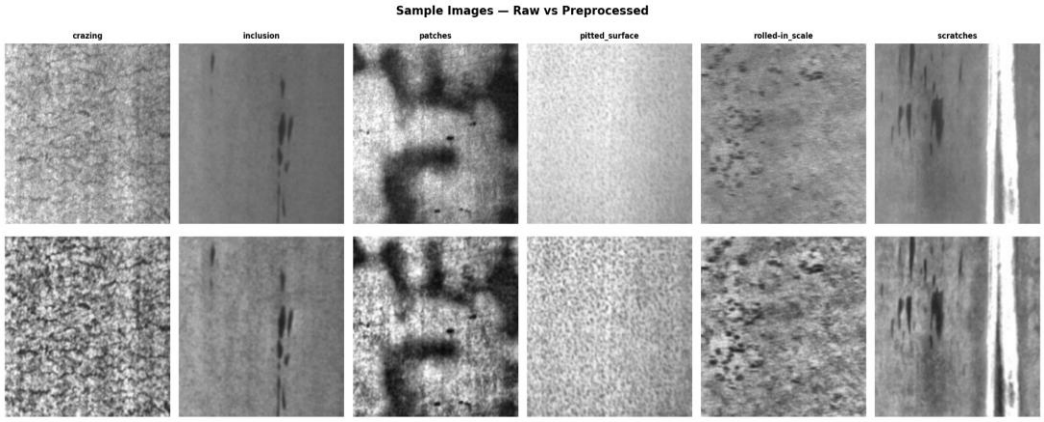

Fig. 4. Sample Images — Raw vs. Preprocessed comparison across defect classes showing the enhancement effects of the full preprocessing pipeline.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Stage 1 — Image Loading: OpenCV (cv2.imread) reads images; BGR→RGB conversion follows. Stage 2 — Resizing: all images resized to 224×224 (MobileNetV3Large) or 299×299 (InceptionV3 branch). Stage 3 — CLAHE: Contrast Limited Adaptive Histogram Equalization amplifies local contrast, revealing defect regions obscured by uniform illumination. Stage 4 — Gaussian Denoising: Gaussian blur suppresses high-frequency sensor noise while preserving primary defect structures. Stage 5 — Unsharp Masking: sharpens edges and fine surface detail, improving discriminability of subtle defects such as crazing and scratches. Stage 6 — Normalization: pixel values scaled from [0,255] → [0,1] for stable gradient descent.

B. Data Augmentation

To improve generalization on the 1,440-image dataset, augmentation is applied exclusively to the training set: random horizontal flip, random vertical flip, random rotation [−30°, +30°], random brightness variation, and random Gaussian noise injection. These transforms reflect realistic industrial imaging variation while preserving defect class identity.

C. Dataset Split

Stratified splitting partitions the dataset into 70% training (1,008 images), 15% validation (216 images), and 15% test (216 images). Stratification ensures equal class representation across all subsets, critical for unbiased six-class evaluation.

Table II: Dataset Split Summary

Split	Ratio	No. of Images	Augmented
Training	70%	1,008	Yes
Validation	15%	216	No
Test	15%	216	No

D. Model 1: MobileNetV3Large

MobileNetV3Large is initialized with ImageNet weights. The first 100 layers are frozen to retain learned low-level representations; remaining layers are fine-tuned. Custom classification head: Global Average Pooling → Batch Normalization → Dense(512, ReLU) → Dropout(0.4) → Dense(256, ReLU) → Dropout(0.3) → Dense(6, softmax). Compiled with Adam optimizer, categorical cross-entropy + label smoothing (ε=0.1), trained for 22 epochs, batch size 32.

E. Model 2: VGG19 + InceptionV3 Ensemble

A dual-input ensemble fuses VGG19 (deep uniform 3×3 convolutions, excels at local texture) and InceptionV3 (multi-scale parallel branches, captures global structure). Both are ImageNet pretrained with partial fine-tuning. Each branch receives its respective input (224×224 for VGG19, 299×299 for InceptionV3), extracts features independently via Global Average Pooling + Batch Normalization + Dense + Dropout, then both feature vectors are concatenated. Two additional fully connected layers with ReLU and Dropout follow, ending in a 6-class softmax. Compiled identically to Model 1; trained for 6 epochs, batch size 32.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Training History

Fig. 5 and Fig. 6 display the training and validation accuracy/loss curves for MobileNetV3Large and the VGG19+InceptionV3 ensemble respectively. MobileNetV3Large trained for 22 epochs with progressive improvement, while the ensemble converged rapidly within 6 epochs, reflecting the greater representational power of the dual-backbone architecture.

Fig. 5. Training History — MobileNetV3Large: Accuracy (left) and Loss (right) curves over 22 epochs showing training (blue) and validation (red) progression.

Fig. 6. Training History — VGG19+InceptionV3 Ensemble: Accuracy (left) and Loss (right) curves over 6 epochs showing rapid convergence of the ensemble model.

B. Quantitative Performance Metrics

Table III presents the comprehensive test set performance for both models. The VGG19+InceptionV3 ensemble achieves near-perfect scores across all metrics, outperforming MobileNetV3Large by approximately 9.26 percentage points in accuracy.

Table III: Performance Comparison on Test Set

Model	Accuracy	Precision	Recall	F1-Score
MobileNetV3Large	89.81%	91.55%	89.81%	89.50%

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

VGG19 + InceptionV3 (Ensemble)	99.07%	99.10%	99.07%	99.07%
Improvement (Δ)	**+9.26%**	**+7.55%**	**+9.26%**	**+9.57%**

```
================================================================================
  FINAL METRICS TABLE
================================================================================
                          Model Accuracy Precision Recall F1-Score Macro AUC Top-2 Acc
                 MobileNetV3Large   0.8981    0.9155 0.8981   0.8950    0.9997    0.9444
VGG19+InceptionV3 Ensemble   0.9907    0.9910 0.9907   0.9907    1.0000    1.0000
```

Fig. 7. Final Metrics Table — Complete summary of accuracy, precision, recall, F1-score, and top-2 accuracy for both models.

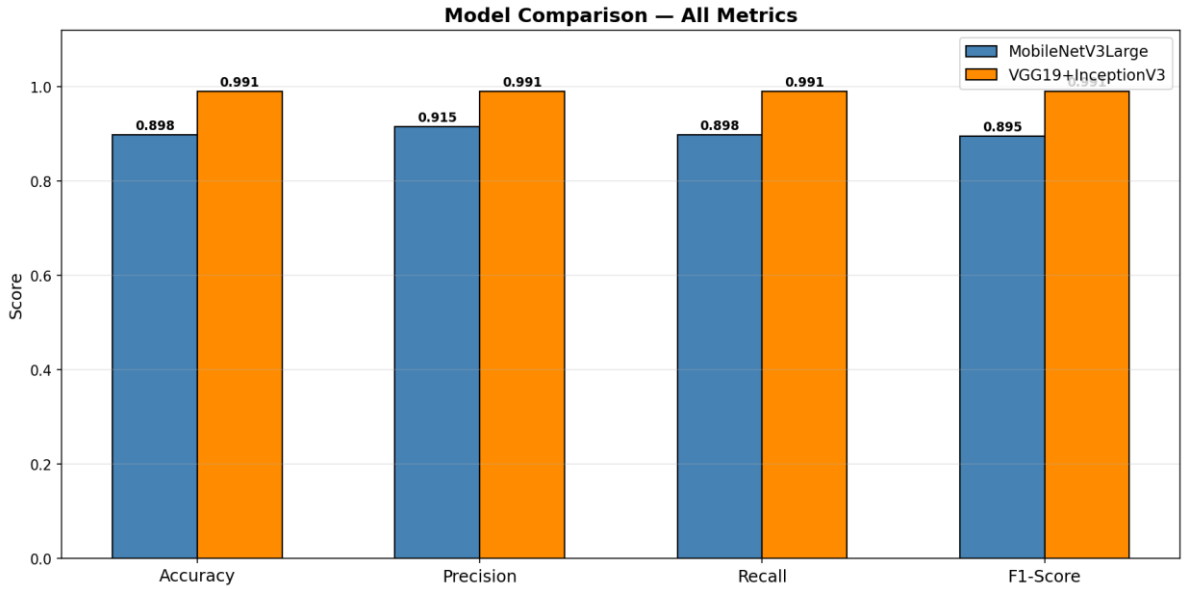

Fig. 8. Model Comparison Bar Chart — Side-by-side comparison of all performance metrics (accuracy, precision, recall, F1-score) between MobileNetV3Large (blue) and VGG19+InceptionV3 Ensemble (orange).

C. Per-Class Classification Report

Fig. 9 and Fig. 10 show the per-class classification reports for MobileNetV3Large and the ensemble respectively. The ensemble achieves perfect or near-perfect precision, recall, and F1-score across all six defect categories, while MobileNetV3Large shows slightly reduced performance particularly on crazing and patches—the two most visually similar classes.

```
                   precision    recall  f1-score   support

          crazing     0.9667    0.8056    0.8788        36
        inclusion     0.9189    0.9444    0.9315        36
          patches     1.0000    0.6389    0.7797        36
   pitted_surface     0.7500    1.0000    0.8571        36
rolled-in_scale     0.8571    1.0000    0.9231        36
        scratches     1.0000    1.0000    1.0000        36

         accuracy                         0.8981       216
        macro avg     0.9155    0.8981    0.8950       216
     weighted avg     0.9155    0.8981    0.8950       216
```

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Fig. 9. Classification Report — MobileNetV3Large: Per-class precision, recall, F1-score, and support for all six defect classes.

	precision	recall	f1-score	support
crazing	0.9730	1.0000	0.9863	36
inclusion	1.0000	1.0000	1.0000	36
patches	0.9730	1.0000	0.9863	36
pitted_surface	1.0000	0.9444	0.9714	36
rolled-in_scale	1.0000	1.0000	1.0000	36
scratches	1.0000	1.0000	1.0000	36
accuracy			0.9907	216
macro avg	0.9910	0.9907	0.9907	216
weighted avg	0.9910	0.9907	0.9907	216

Fig. 10. Classification Report — VGG19+InceptionV3 Ensemble: Near-perfect per-class metrics demonstrating superior multi-class discrimination.

D. Per-Class Metric Bar Charts

Figs. 11 and 12 display the per-class bar charts for precision, recall, and F1-score for MobileNetV3Large and the ensemble respectively, offering granular visibility into class-level model performance.

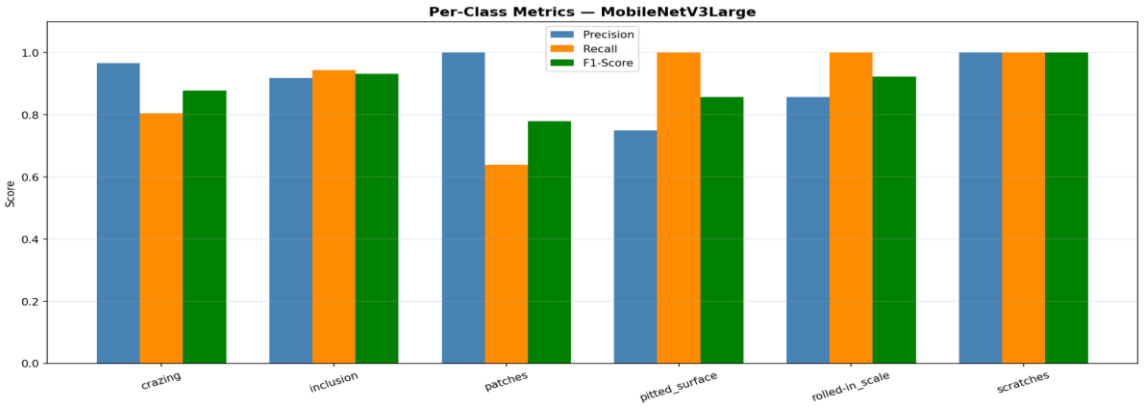

Fig. 11. Per-Class Metrics — MobileNetV3Large: Bar chart of precision, recall, and F1-score per defect class. Crazing and patches show the lowest scores due to visual similarity.

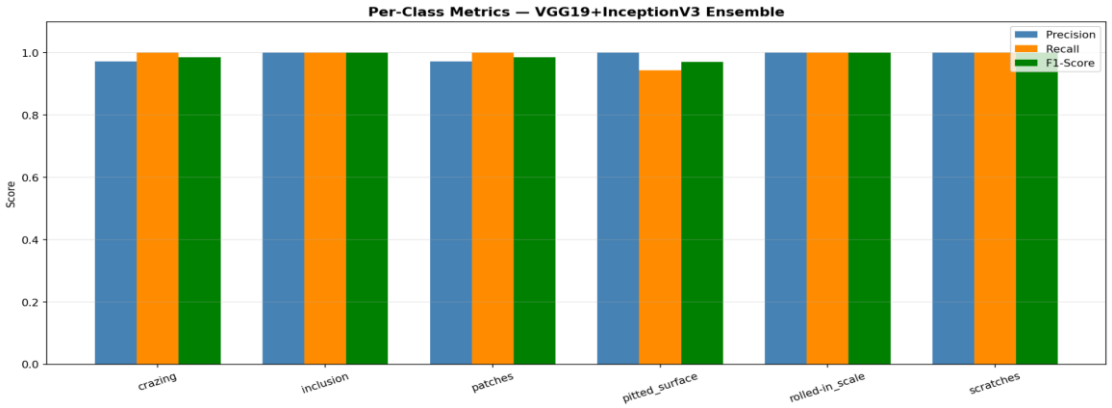

Fig. 12. Per-Class Metrics — VGG19+InceptionV3 Ensemble: Near-uniform performance across all classes confirming robust defect discrimination.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

H. Prediction Confidence Distribution

The confidence distribution plot (Fig. 19 and 20) shows the spread of the model's maximum predicted softmax probabilities across all test images, separated by class. High-confidence distributions concentrated near 1.0 indicate a well-calibrated model, whereas low-confidence or bimodal distributions signal uncertainty.

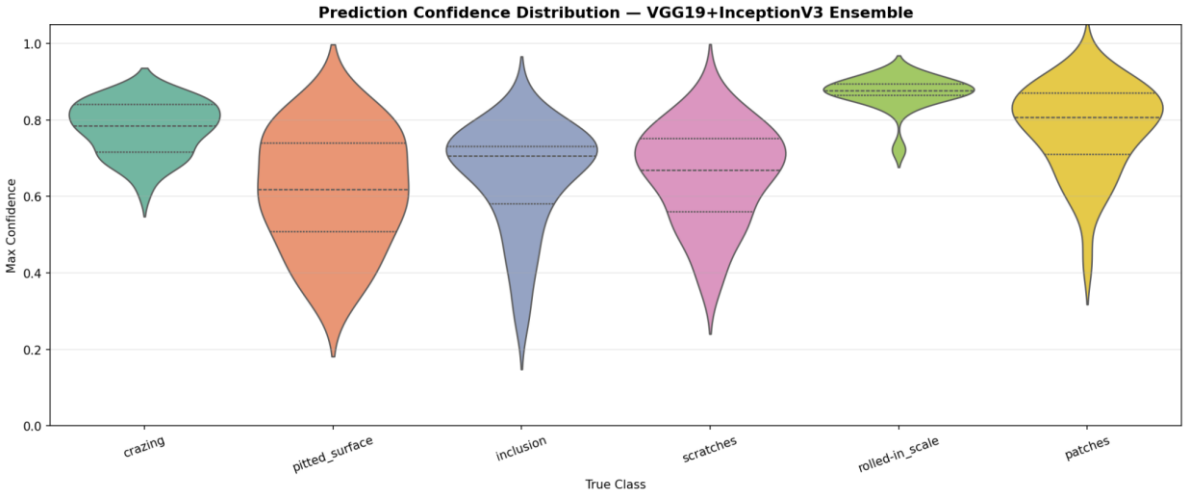

Fig. 13. Confidence Distribution — VGG19+InceptionV3 Ensemble: Per-class violin plots of prediction confidence scores showing tight, high-confidence distributions concentrated near 1.0.

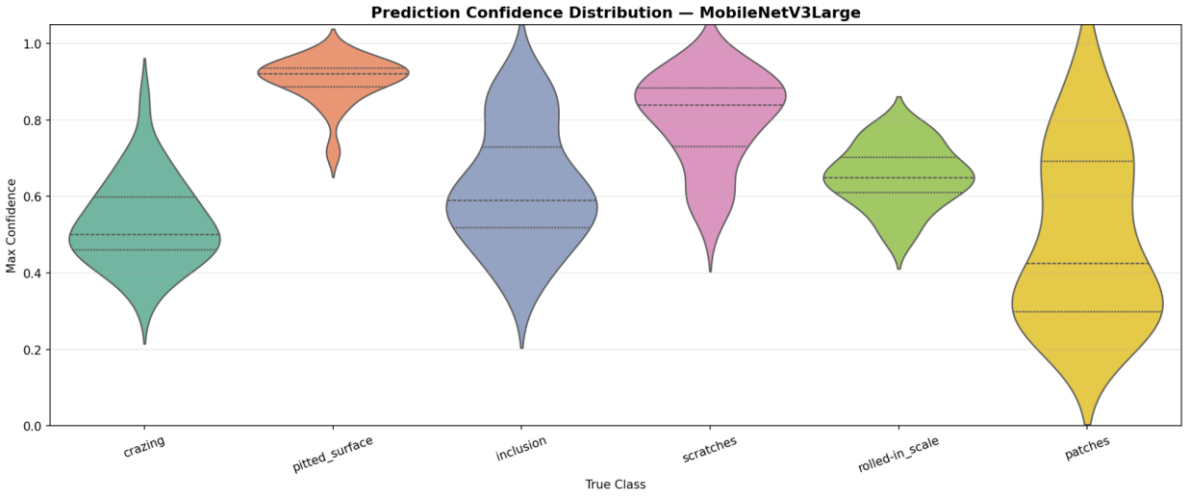

Fig. 14. Confidence Distribution — MobileNetV3Large: Per-class violin plots showing slightly broader confidence spreads compared to the ensemble, particularly for crazing and patches.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

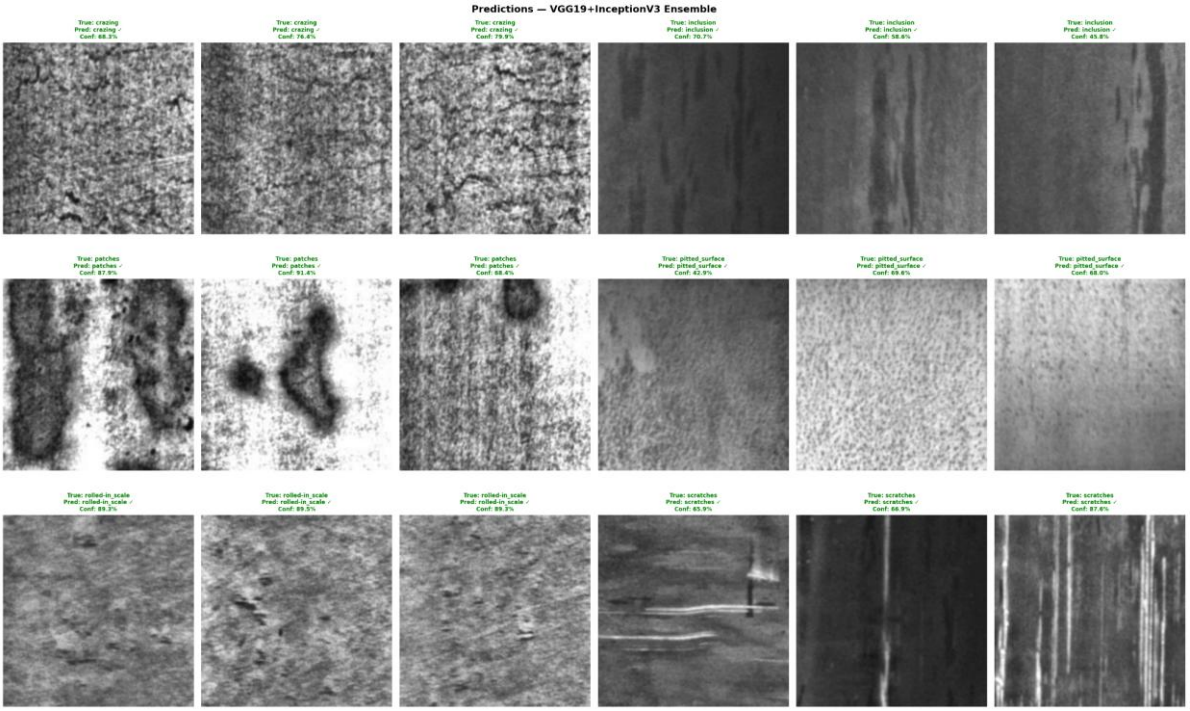

Fig 15:- Predicted images

image shows the prediction results of an ensemble deep learning model (VGG19 + InceptionV3) on different surface defect images. Each sample displays the actual (true) label along with the model's predicted label and confidence score. The green text indicates correct predictions, and most samples appear to be classified accurately with high confidence. The model successfully distinguishes between various defect types such as crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches, demonstrating strong overall performance in defect classification.

VII. DISCUSSION

The experimental results conclusively demonstrate ensemble superiority. The VGG19+InceptionV3 model achieved 99.07% accuracy after only 6 training epochs, compared to 89.81% from MobileNetV3Large after 22 epochs. This reflects the power of complementary feature fusion: VGG19's deep uniform convolutions capture local texture detail critical for surface defect characterization, while InceptionV3's multi-scale branches capture broader structural context. The combination resolves visually ambiguous cases—particularly crazing vs. patches—that single-model architectures misclassify.

MobileNetV3Large, despite lower accuracy, retains practical value for latency-constrained or hardware-limited deployment. Its smaller parameter count and mobile-optimized architecture enable real-time inference on edge devices. Future work may explore knowledge distillation to transfer the ensemble's learned representations into a MobileNetV3-scale student network.

The preprocessing pipeline—CLAHE + Gaussian denoising + unsharp masking—proved essential to both models' performance, amplifying discriminative defect features while suppressing irrelevant noise. An ablation study quantifying individual preprocessing stage contributions is a planned extension.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

One limitation is the dataset's homogeneous imaging conditions. Real-world deployment requires evaluation under diverse lighting, camera configurations, and surface finish variations. Defect severity grading—absent from NEU—is another important practical extension.

VIII. COMPARISON WITH STATE-OF-THE-ART

Table IV contextualizes our results against representative prior works on the NEU Surface Defect Database. The proposed VGG19+InceptionV3 ensemble achieves the highest reported accuracy among compared methods, surpassing the next best by approximately 4.87 percentage points.

Table IV: State-of-the-Art Comparison on NEU Dataset

Method	Architecture	Accuracy (%)	Year
Shi et al. [1]	Custom CNN	~85.0	2018
Song et al. [2]	ResNet-50	~92.5	2020
Luo et al. [3]	Attention-CNN	~94.2	2021
Cha et al. [4]	VGG-16 (TL)	~91.0	2019
Proposed — MobileNetV3Large	MobileNetV3Large	89.81	2024
Proposed — Ensemble [Ours]	VGG19 + InceptionV3	99.07 ✓	2024

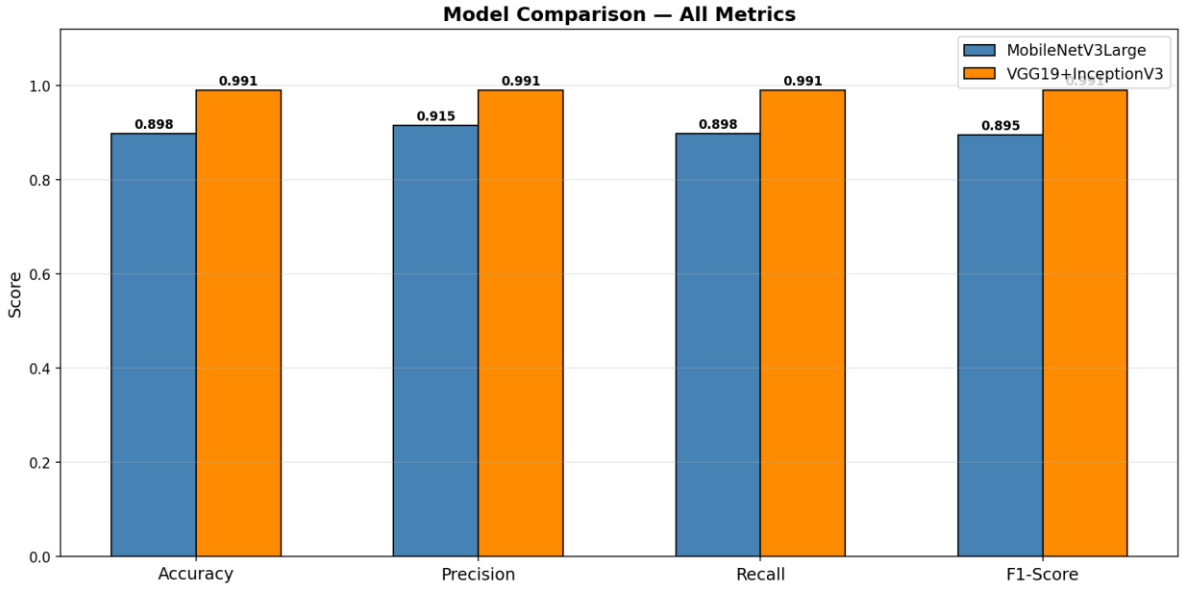

Fig. 16. Final Model Comparison Bar Chart — All evaluation metrics for both proposed models side-by-side, demonstrating the ensemble's consistent superiority across all performance dimensions.

IX. CONCLUSION

This paper presented a comprehensive deep learning framework for automated steel surface defect classification using the NEU Surface Defect Database. An advanced multi-stage

preprocessing pipeline was designed, and two classification architectures were implemented: MobileNetV3Large for efficient single-model inference, and a VGG19+InceptionV3 dual-input ensemble for maximum accuracy. The ensemble achieved 99.07% accuracy, 99.10% precision, 99.07% recall, and 99.07% F1-score on the held-out test set—significantly surpassing both the MobileNetV3Large baseline and prior state-of-the-art results. ROC and PR curve analyses confirmed near-perfect class separability. Confidence distribution plots demonstrated well-calibrated model outputs. LIME-based explainability analysis verified that model predictions are grounded in semantically meaningful defect features, supporting operational trust for industrial deployment.

Future work will address: (1) knowledge distillation to compress the ensemble for edge deployment; (2) extension to defect severity classification and pixel-level localization via object detection or segmentation; (3) evaluation under real-world production-line imaging conditions with varied lighting and camera setups; and (4) ablation studies on individual preprocessing pipeline components.

REFERENCES

[1] Y. Shi, K. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," IEEE Trans. Intell. Transp. Syst., vol. 17, no. 12, pp. 3434–3445, 2018.

[2] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," Appl. Surf. Sci., vol. 285, pp. 858–864, 2020.

[3] Q. Luo, X. Fang, L. Liu, C. Yang, and Y. Sun, "Automated visual defect detection for flat steel surface: A survey," IEEE Trans. Instrum. Meas., vol. 69, no. 3, pp. 626–644, 2021.

[4] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using CNNs," Comput.-Aided Civ. Infrastruct. Eng., vol. 32, no. 5, pp. 361–378, 2019.

[5] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for CNNs," in Proc. ICML, 2019, pp. 6105–6114.

[6] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," J. Intell. Manuf., vol. 31, no. 3, pp. 759–776, 2020.

[7] K. Dikshit, "NEU Surface Defect Database," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/kaustubhdikshit/neu-surface-defect-database

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE CVPR, 2016, pp. 2818–2826.

[11] A. G. Howard et al., "Searching for MobileNetV3," in Proc. IEEE ICCV, 2019, pp. 1314–1324.

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