Advances in Consumer Research
https://acr-journal.com/
Volume-2 | Issue-6 | Dec 2025

Original Researcher Article

Detection of Steel Manufacturing Defects Using Few Shot Learning Multimodal Model

Rapaka Praveen[1], K. Ch Apparao[2]*, Chinta Sandeep[3] and GVR Seshagiri Rao[4]
[1,2,3,4]Department of Mechanical Engineering, Institute of Aeronautical Engineering, Hyderabad, India-500043.
***Corresponding author**:
Email:ID: kchapparao@gmail.com

ABSTRACT
Steel has wide range of yield strength with high modulus of elasticity. It is corrosion-resistant and long-lasting, making thinner and more durable structures. These properties make steel versatile and efficient material in construction and automotive. It enables the creation of innovative designs that are both strong and economical, while its durability ensures long-term performance. In some cases, defects in steel will occur which show impact on the final products. Figuring out these defects requires rigorous quality control during manufacturing and proper handling during fabrication and assembly. In this project, steel defects are detected using multimodal. As multimodal has the capability to capture insights from limited data, it can be used in the industries to find out the defects
Keywords: Steel; Construction; Automotive; Defects; Multimodal

INTRODUCTION:

Steel provides a unique combination of strength, ductility, and corrosion resistance, making it an essential material for modern industrial applications. It is essential in sectors including energy production, construction, automobile production, and aerospace because to its affordability and adaptability. Steel makes a substantial contribution to infrastructure development and improvements in technology by satisfying the demanding criteria of many applications [1].

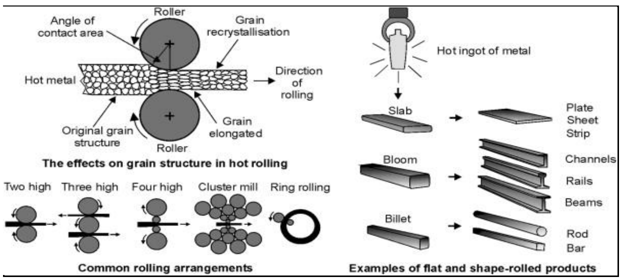

Fig 1.1. Hot Rolling Process [3]

Steel needs to go through regulated manufacturing procedures in order meet the demands of these industries. One essential production forming technique that turns semi-finished steel materials into useful shapes including sheets, plates, bars, and structural sections is hot rolling [2]. One of the most important processes in the production of steel after casting is hot rolling, which is a fundamental metal forming process. Hot rolling is essentially the process of reducing the cross-sectional area of heated steel and shaping it into the appropriate shapes by passing it through a number of spinning rolls. Hot rolling offers clear advantages over cold working since it is carried out above the steel's recrystallization temperature, which is normally between 1100 and 1300 °C. Steel's ductility increases, and its yield strength drastically decreases at such high temperatures, enabling significant plastic deformations. Accordingly, cast ingots, slabs, or billets can be efficiently broken down into plates, sheets, bars, and other forms by achieving significant thickness reductions in one or a few passes [4, 5].

In a conventional hot rolling mill, the process starts with warming the steel input (slab, bloom, or billet) in a furnace to the desired rolling temperature. Once evenly heated, the red-hot steel is descaling, in which high-pressure water jets remove the iron oxide scale that generated in the furnace. Next, the steel is sent through a roughing mill, which is a collection of rolls that forcefully crush and elongate the piece, significantly lowering thickness in preparation for final rolling. To achieve an intermediate thickness, the roughing stage may need many passes, often with a reversing mill that feeds the piece back and forth. The work then moves to the finishing mill, which is a series of finer roll stands that gradually bring the steel to its final dimensions and surface finish. Throughout these steps, the steel's temperature is carefully maintained to stay within the hot-work range. After the last rolling pass, the product is cooled, usually with regulated water sprays on a run-out table, because cooling pace might affect mechanical qualities. Finally, depending on the product type, the rolled steel is coiled or cut to length before export. This hot rolling sequence is typically a highly automated, continuous flow in modern mills, with computers coordinating roll speeds, gaps, and temperatures to ensure consistent results. [6]

Steel plays an important role in assuring passenger safety, structural integrity, and energy efficiency in automobiles. Advanced high-strength steel (AHSS) is commonly utilized in vehicle designs to minimize weight while maintaining durability. This lightweight design is critical for increasing fuel efficiency, fulfilling severe emissions rules, and maintaining high crashworthiness standards. Steel's high energy absorption ability contributes significantly to vehicle safety during crashes. In the case of a collision, steel components bend in a controlled manner to dissipate energy, protecting passengers by reducing the force delivered into the cabin. Furthermore, steel's versatility enables manufacturers to create sophisticated forms and components, such as reinforced pillars, crumple zones, and high-strength chassis sections, which are required in current vehicle designs.

The demand for electric vehicles (EVs) has boosted the value of steel. For example, steel is utilized in EV battery casings to protect them from mechanical harm while also providing thermal management features. Steel's dual purpose makes it essential for achieving the stringent safety and performance demands of EVs.

Steel plays an equally important function in building, serving as the foundation for modern architecture and infrastructure. Its high strength-to-weight ratio and capacity to survive harsh weather conditions make it an ideal material for skyscrapers, bridges, and industrial structures. Engineers and architects use steel to develop complicated structures that require both strength and flexibility, allowing for unique designs that push the frontiers of construction. One of steel's primary advantages is its resistance to high temperatures and heavy loads, making it perfect for vital infrastructure projects. In addition to its strength, steel's longevity guarantees that constructions may last for decades with little upkeep. For example, the use of weathering steel in bridges and outdoor constructions improves corrosion resistance, greatly increasing their lifespan under hard conditions.

Sustainability has been a prominent priority in building, and steel's great recyclability helps to promote environmentally friendly methods. Recycled steel keeps its qualities and may be reused several times, lowering the environmental impact of construction projects. The steel sector has also made progress in implementing green technology, such as low-carbon production processes and renewable energy sources, which aligns with global sustainability goals.

Despite its numerous benefits, steel flaws may have major ramifications across sectors, notably in terms of safety, performance, and cost. Defects, whether during manufacturing or handling, jeopardize the integrity of steel components and impair their efficacy in key applications. Steel flaws in the automobile industry pose substantial concerns to passenger safety and vehicle performance. Microcracks or voids inside steel components can spread under stress, potentially resulting in unexpected part failure during operation. For example, a crack in a suspension or chassis component might cause the vehicle to lose control, endangering occupants and other road users.

Aside from safety considerations, faults can impair vehicle performance and dependability. Corrosion caused by surface flaws or inclusions can shorten the life of components, requiring frequent replacements and raising maintenance expenses. Defects in production can result in costly recalls, brand reputation harm, and a loss of consumer confidence. Steel flaws in EVs can impair the safety and performance of battery enclosures, affecting the protection of energy storage systems. Given the high stakes in EV safety, defect identification is important for ensuring compliance with industry requirements.

Steel flaws may have disastrous repercussions in the building industry. Structural steel is used to carry enormous loads and stabilize buildings, bridges, and other structures. Cracks, inclusions, and poor welds limit steel components' load-bearing capability, increasing the likelihood of structural collapse.

For example, a crack in a steel beam used in a high-rise building might jeopardize the structural integrity of the entire structure. Similarly, inclusions or cavities in steel used in bridge construction can cause premature fatigue and eventually collapse under repetitive load. Such failures not only cause large financial losses but also pose serious threats to human life and property. Undetected steel faults have an influence on maintenance costs in building projects. Small faults might compound over time, necessitating periodic inspections, repairs, or even replacements. This increases the operating expenses of maintaining infrastructure and diminishes its overall efficiency.

Cost of Quality (CoQ) is a way for assessing the expenses that businesses pay in ensuring that products satisfy quality requirements, as well as the costs of manufacturing things that do not meet quality standards.

COQ is classified into two categories:

- Cost of Good Quality (COGQ), and
- Cost of Poor Quality (COPQ).

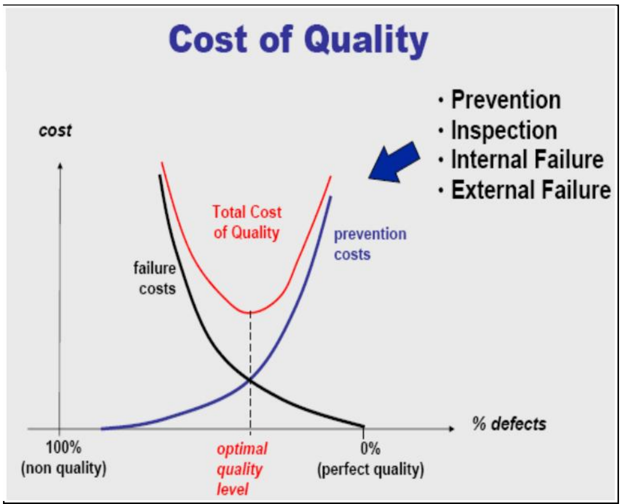

Fig 1.2. Cost of Quality Graphical Representation [8]

The COQG is classified into two components:

- prevention cost, and
- appraisal cost.

Prevention costs are connected with actions geared expressly to prevent poor product quality, such as quality planning, quality audits, and market research.

How to cite : Rapaka Praveen, K. Ch Apparao, Chinta Sandeep and GVR Seshagiri Rao, Detection of Steel Manufacturing Defects Using Few Shot Learning Multimodal Model *Advances in Consumer Research*. 2025;2(6): 3118-3125

Appraisal costs are connected when an organization pays for an expert to identify poor product quality through supplier inspection and lab testing.

The COPQ is classified into two components:

- internal failures, and
- external failures.

Internal failure costs result from a product's failure to meet needed quality requirements.

External failures cost money because buyers reject things after they are delivered.

LITERATURE REVIEW

Tewogbade Shakir Adeyemi designed a deep learning system that uses DenseNet121 for defect classification and DeepLabV3 for segmentation. DenseNet121 obtained an accuracy of 92.34%, outperforming benchmarks like VGG16 (72.59%) and ResNet50 (92.01%). DeepLabV3 performed well in defect localization, with a Dice coefficient of 84.21% during training and 69.77% during validation, notably in recognizing Defect 4 (87.69%) but falling short on Defect 1 (64.81%). The system obtained 92.31% precision by combining classification, segmentation, and thresholding approaches, outperforming previous models in terms of accuracy and reliability [1]. Author suggested the approach of multimodal learning to create more robust and reliable defect detection ML system [1].

Abu, M., A. Amir et al., performed steel defects detection using transfer learning on SEVERSTAL dataset to classify the defective and non-defective image and used NEU dataset to classify ‘rolled in scale’‘patches’, ‘crazing’, ’pitted surface’, ‘inclusion’ and ‘scratches’. MobileNet architecture performed better than ResNet, VGG, and DenseNet for binary classification and multi-class classification [2]. In Future, to improve the performance of the steel defect detection authors planned for incorporating MobileNet model as feature map and combine it with custom layer of convolutional neural network.

Damacharla, Praveen et al., proposed Transfer Learning-based U-Net (TLU-Net) framework for steel surface defect detection. U-Net architecture was used as the base and explore ResNet and DenseNet. Comparison on performance was executed using random initialization and the pre-trained networks trained using the ImageNet data set. As a result, transfer learning performed better than random initialization. As it was difficult for transfer learning to handle more complex shapes and new rare defects, authors wanted to explore the semi/weakly supervised learning approaches to reduce the annotated training data requirement [3].

Boikov, Aleksei et al., used U-Net and Xception architectures on synthetic dataset of steel defects which was generated using the 3D graphics editor Blender. Then these two models were tested on Severstal: Steel Defect Detection dataset. U-Net achieved a dice score of 0.56 on real data and 0.63 on syntenic data where as Xception achieved Precision and Recall more than 0.87 on real data and more than 0.8 on syntenic data [4].

Qian, Kun utilized U-Net and feature pyramid network (FPN) for real-time semantic segmentation of steel defects. The results show that the adopted architectures have a decent performance regarding both time cost and segmentation accuracy. They achieved dice coefficients over 0.915 and 0.905 at a speed of over 1.5 images per second on the public test set and private test set (Severstal: Steel Defect Detection dataset) on the Kaggle platform, respectively [5].

Chigateri, Keerthana B et al., created models to detect steel defects using SEVERSTAL dataset. These models included binary classification, multiclass classification and U-Net. Binary and multi-label classifications was computed first and then segment images using segmentation models and predict encoded pixels after testing thresholds for each defect [6].

Akhyar, Fityanul et al., proposed a novel deep learning–based surface defect inspection system called the forceful steel defect detector (FDD). This FDD system has random scaling in the data preprocessing pipeline for training and ultimate scaling for testing and combines them with an improved cascade R-CNN detector which involves deformable convolution, deformable RoI pooling, and a guided anchoring RPN [7].

METHODOLOGY

Traditional supervised learning algorithms rely on vast volumes of labelled data to attain good performance. Collecting large, annotated datasets is expensive in many real-world applications, including medical imaging, satellite images, rare defect identification in manufacturing, and low-resource languages in natural language processing. Few-Shot Learning (FSL) overcomes this restriction by allowing models to generalize from a small number of labelled samples per class [9–11].

The central challenge of FSL lies in balancing two competing requirements:

(i) the ability to quickly adapt to unseen classes using very limited data

(ii) the prevention of overfitting to those few examples [12].

Meta learning and non-meta learning are the two approaches to execute FSL as shown in Fig 3.1.

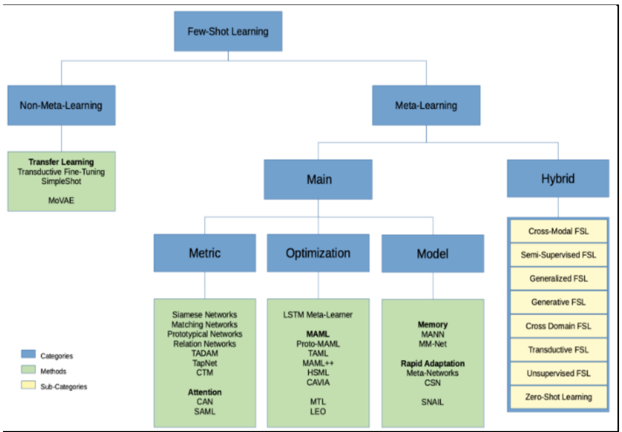

Fig 3.1. FSL Approaches [9]

3.1 Model Design and Architecture

Snell, Swersky, and Zemel [13] established the Prototypical Network, which has become one of the most important techniques to few-shot learning (FSL).

Prototypical networks are a form of machine learning model built for few-shot learning, which involves classifying new categories with only a few labelled samples. Instead, then developing a complicated classifier for each new class, prototypical networks represent each class with a prototype, which is a single vector summarizing the class. This prototype is calculated as the average of the feature vectors in the support set for that class.

When a query is received, the network compares it to each class prototype by measuring the Euclidean distance in the feature space. The query is identified as the class with the closest prototype. This strategy significantly lowers complexity and enables learning with less data.

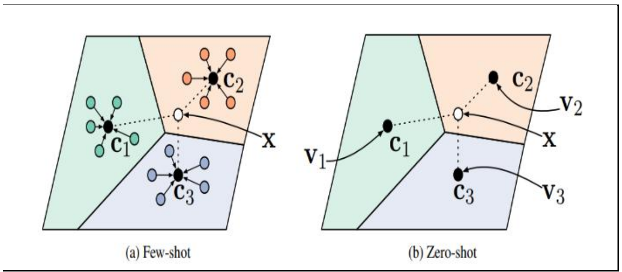

Fig 3.2. Prototypical networks based few-shot and zero-shot [13]

Algorithm 1 Training episode loss computation for prototypical networks. N is the number of examples in the training set, K is the number of classes in the training set, $N_C \leq K$ is the number of classes per episode, N_S is the number of support examples per class, N_Q is the number of query examples per class. RANDOMSAMPLE(S, N) denotes a set of N elements chosen uniformly at random from set S, without replacement.

Input: Training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where each $y_i \in \{1, \ldots, K\}$. $\mathcal{D}_k$ denotes the subset of $\mathcal{D}$ containing all elements $(\mathbf{x}_i, y_i)$ such that $y_i = k$.
Output: The loss J for a randomly generated training episode.

$V \leftarrow$ RANDOMSAMPLE$(\{1, \ldots, K\}, N_C)$ ▷ Select class indices for episode
for k in $\{1, \ldots, N_C\}$ **do**
  $S_k \leftarrow$ RANDOMSAMPLE$(\mathcal{D}_{V_k}, N_S)$ ▷ Select support examples
  $Q_k \leftarrow$ RANDOMSAMPLE$(\mathcal{D}_{V_k} \setminus S_k, N_Q)$ ▷ Select query examples
  $\mathbf{c}_k \leftarrow \frac{1}{N_C} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i)$ ▷ Compute prototype from support examples
end for
$J \leftarrow 0$ ▷ Initialize loss
for k in $\{1, \ldots, N_C\}$ **do**
  for $(\mathbf{x}, y)$ in Q_k **do**
    $J \leftarrow J + \frac{1}{N_C N_Q} \left[d(f_\phi(\mathbf{x}), \mathbf{c}_k)) + \log \sum_{k'} \exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_k)) \right]$ ▷ Update loss
  end for
end for

Fig 3.3. Pseudocode for Prototypical networks [13]

In recent years, deep learning models have achieved substantial advances in computer vision and natural language processing (NLP). However, most classical models are trained in separate tasks such as image classification and text classification using large, labelled datasets. This reduces their capacity to generalize across modalities. Radford et al. [14] proposed Contrastive Language-Image Pretraining (CLIP), a ground-breaking paradigm for learning combined representations of both words and images. The model can interpret visual concepts in natural language without requiring task-specific labelled data.

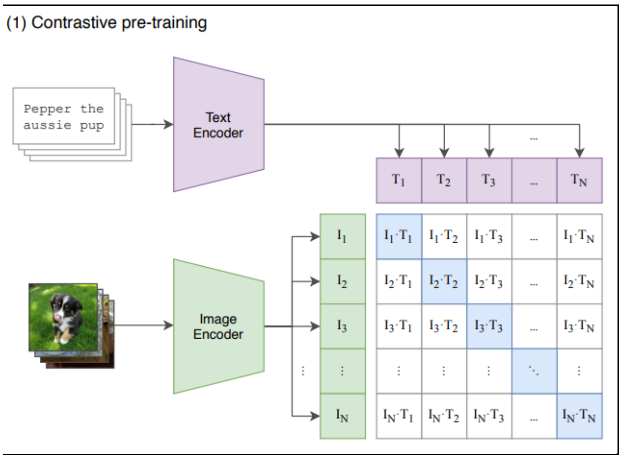

Fig 3.4.4 CLIP approach summary-1 [14]

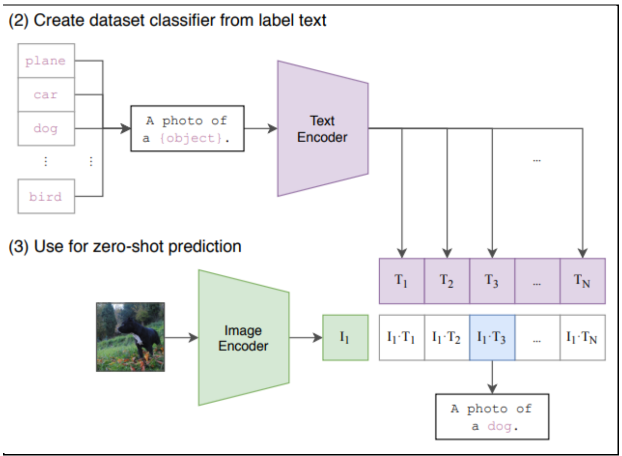

Fig 3.5 CLIP approach summary-2 [14]

CLIP's fundamental idea is to align text and image information in a single embedding space with contrastive learning.

- An image encoder processes the image.
- Transformer-based text encoder handles captions.
- Both images and captions are mapped into the same vector space.
- The model is trained so that the embedding of an image is close to its matching caption and far from unrelated captions.

The training uses a contrastive loss function (InfoNCE) across many image–text pairs.

CLIP can identify new categories without further training. It was trained using 400 million image-text pairs from the internet named WebImage Text, making it more generic than models based on smaller curated datasets. It integrates language with vision, allowing for activities such as image search from text and text-based picture classification.

3.2 Dataset Overview

In this work, NEU Surface Defect Database (NEU-CLS) [2] is utilized. This dataset has 1,800 hot steel surface defect images in total and contains six defect types they are

- rolled-in scale,
- patches,
- crazing,
- pitted surface,
- inclusions and
- scratches

with 300 images for each defect. The images are grayscale with 200 X 200 pixels. Grayscale simplifies things, removes colour variation that might not be relevant to the defect itself, but still captures the texture and shape.

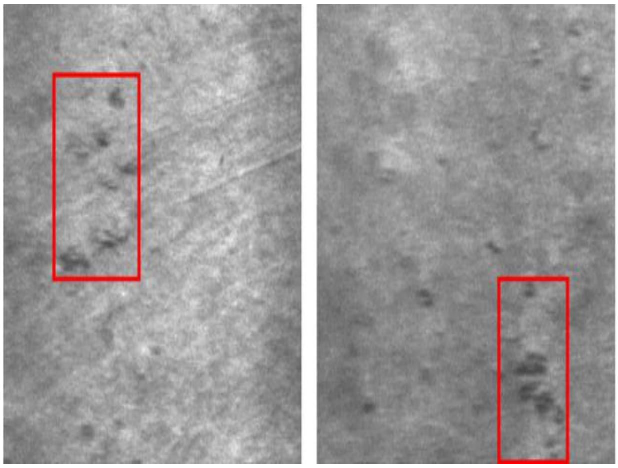

Fig 3.6. Rolled-In Scale Defect

Rolled-In Scale: When steel is heated in a furnace before rolling, it interacts with oxygen, forming a flaky coating known as scale (iron oxide). In the following production process, high-pressure water jets are used to eliminate the scale just before the steel strikes the rollers. If they miss certain locations, brittle scale remains on the surface and travels between the heavy rolls, eventually embedding deep into the steel surface.

Patches: When the slab cools, there is a potential that variable rates of cooling may occur, resulting in unequal oxidation processes. This discrepancy in the environment causes an apparent inconsistency on the surface, i.e. irregular patches where the colour or texture is simply off.

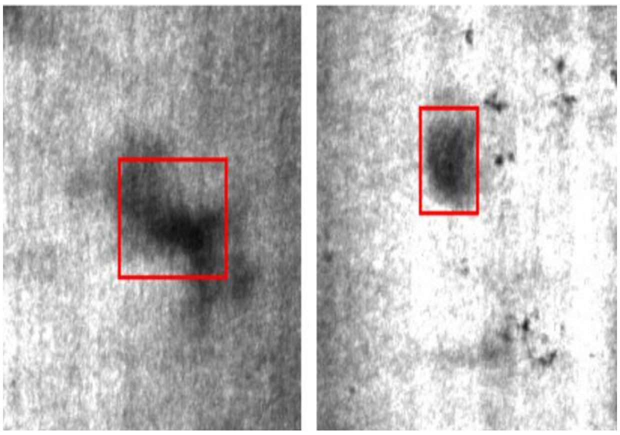

Fig 3.7. Patches Defect

Crazing: There are immense pressures at work when the steel is pressed through the rollers to reduce its thickness and then cools. If the steel is cooled too quickly or reduced too aggressively in one go, the resultant surface loses integrity and develops a thin, web-like pattern of microscopic fractures, similar to a spider web on the steel, caused by mechanical and thermal stress.

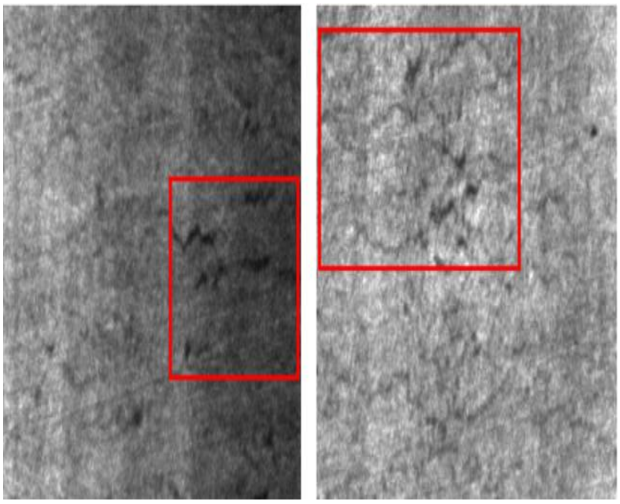

Fig 3.8. Crazing

Pitted Surface: Small cavities or holes on the surface result from the interaction of gases or liquids with heated steel. These gasses can become trapped against the surface when rolling, causing small gaps as the metal hardens around them. During descaling, as water jets strike the heated surface, water vaporizes explosively against the surface, causing localized corrosion in the furnace.

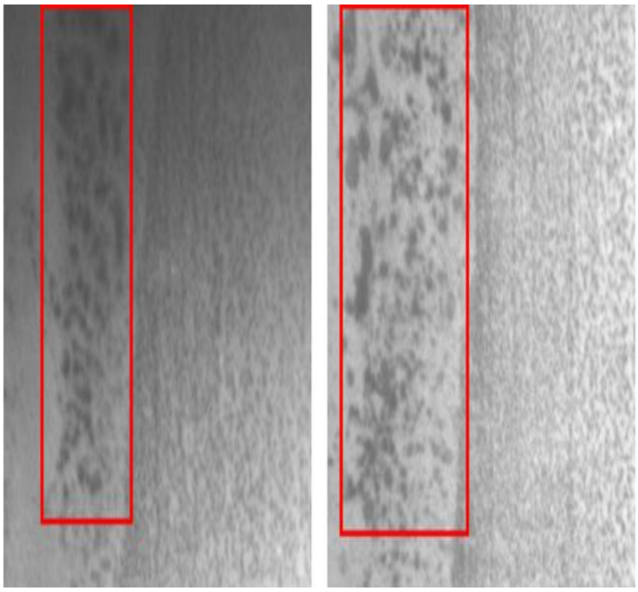

Fig 3.9. Pitted Surface Defect

Inclusions: Unwanted small oxide, sulphides, and silicate particles may become involved in the casting process and are not completely cleaned before solidification. These are non-metallic microscopic particles that become trapped within the steel matrix itself. When the steel is rolled out, the imbedded small particles may become exposed or form stress points near the surface, resulting in those rough hard places.

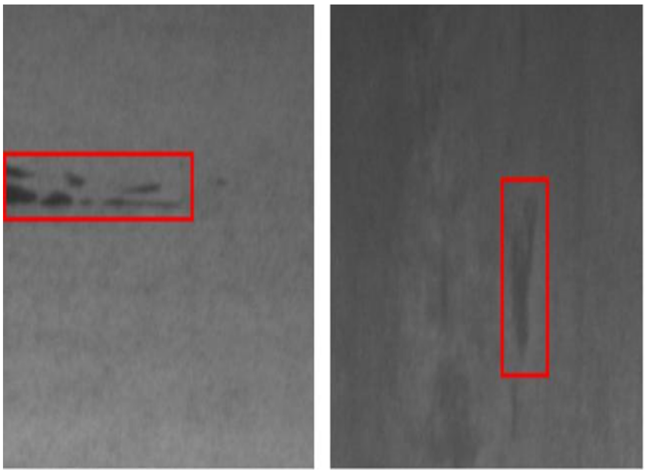

Fig 3.10. Inclusions Defect

Scratches: Misalignment in the rolling mill, or possibly a piece of sharp metal, a burr, being lodged on a guide or a roller, generates mechanical scars, resulting in straight grooves going along the rolling direction.

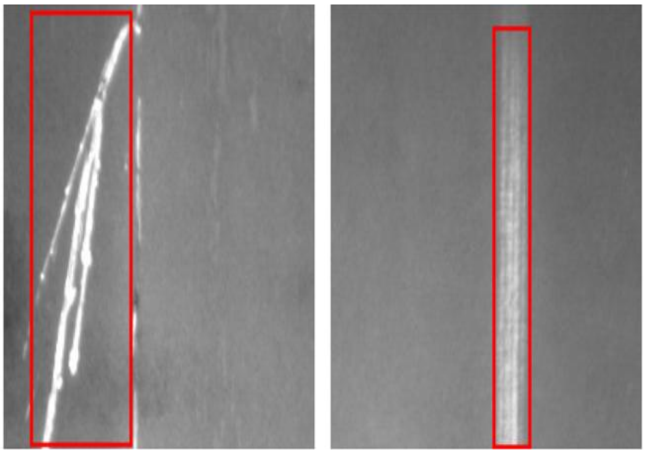

Fig 3.11. Scratches Defect

RESULTS

Performance metrics play a critical role in evaluating the proposed model's efficiency and effectiveness. Accuracy measures the overall correctness of the model in classifying and segmenting defects.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Table 4.1. Experiment with Support Set = 5

Defect Class	Accuracy
Crazing	0.8524
Inclusion	0.4465
Patches	0.9041
Pitted Surface	0.7048
Rolled-in Scale	0.9779
Scratches	0.8487

Table 4.2. Experiment with Support Set = 10

Defect Class	Accuracy
Crazing	0.8985
Inclusion	0.2293
Patches	0.9624
Pitted Surface	0.6504
Rolled-in Scale	0.9662
Scratches	0.8571

Table 4.3. Experiment with Support Set = 15

Defect Class	Accuracy
Crazing	0.8889
Inclusion	0.1188
Patches	0.9923
Pitted Surface	0.5326
Rolled-in Scale	0.9655
Scratches	0.9020

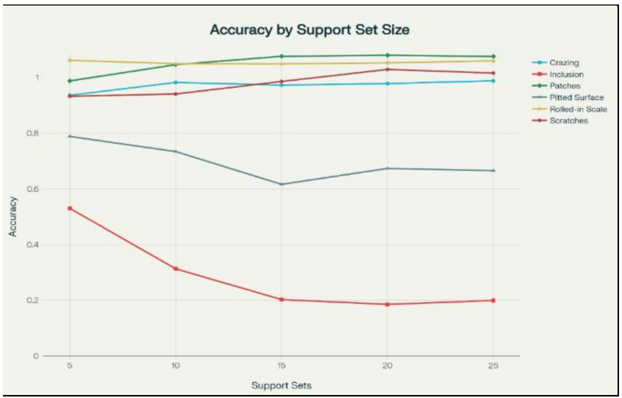

Fig 4.1 Each Class Accuracy Vs Support Set Size

Table 4.4. Experiment with Support Set = 20

Defect Class	Accuracy
Crazing	0.8945
Inclusion	0.1016
Patches	0.9961
Pitted Surface	0.5898
Rolled-in Scale	0.9688
Scratches	0.9453

Table 4.5. Experiment with Support Set = 25

Defect Class	Accuracy
Crazing	0.9044
Inclusion	0.1155
Patches	0.9920
Pitted Surface	0.5817
Rolled-in Scale	0.9761
Scratches	0.9323

Table 4.6. Overall Accuracy with different Support Sets

Support Set	Text-only	Image-only	Fusion
5	0.1722	0.7891	0.7657
10	0.1723	0.7607	0.7751
15	0.1724	0.7344	0.7708
20	0.1719	0.7493	0.8073
25	0.1720	0.7503	0.8161

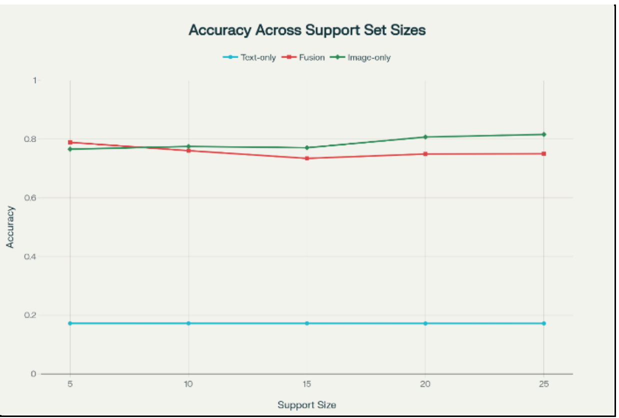

Fig 4.2 Overall Accuracy Vs Support Set Size

CONCLUSIONS

In this work, experimentation was performed using few-shot learning and Mutlimodal model to detect steel defects. According to the results of the experimentation that was conducted, the following conclusions are made:

- Crazing, Patches, Rolled-in Scale, and Scratches consistently achieved high accuracies i.e., greater than 85%, even with a small support set 5 shots.
- Patches and Rolled-in Scale in particular show remarkable robustness, with accuracies above 95% across all support sizes.
- Scratches gradually improve, reaching 94% accuracy approximately at 20 to 25 shots.
- Pitted Surface shows moderate accuracy of 53 to 70%, fluctuating across support sizes. This indicates that this class is more sensitive to prototype quality and requires more discriminative features.
- Inclusion is the most challenging class. Accuracy deteriorates as support size increases (from approximately 45% at 5 shots to 10–12% at 20–25 shots). This suggests significant visual overlap with other defect classes or lack of discriminative CLIP embeddings for this defect type.
- Increasing the support set size from 5 to 25 does not always yield consistent improvements.
- Crazing and Scratches show gradual gains.
- Inclusion suffers from severe degradation, possibly because averaging more visually ambiguous samples dilutes prototype quality.
- Pitted Surface fluctuates, showing instability in prototype representation.
- Text-only accuracy remains near random chance approximately 17% across all experiments. This confirms that CLIP's text embeddings are not directly effective for describing grayscale industrial defects, due to the domain gap between pretraining (natural images, internet captions) and steel surface textures.
- Image-only accuracy ranges between approximately 73–79% at small support sets and improves to 81.6% at 25 shots. This demonstrates that prototypes built from visual embeddings are informative, and accuracy scales modestly with additional labelled examples.

Fusion accuracy (text + image, α=0.5) slightly improves over image-only for most support sizes, peaking at 81.6% for 25 shots. The fusion benefit is marginal because text prototypes contribute weakly; in fact, at low support sizes (5 shots), fusion slightly underperforms image-only.

.

REFERENCES

1. Adeyemi, Tewogbade Shakir. "Defect Detection in Manufacturing: An Integrated Deep Learning Approach." Journal of Computer and Communications 12, no. 10 (2024): 153-176.
2. Abu, M., A. Amir, Y. H. Lean, N. and S. A. Azemi. "The performance analysis of transfer learning for steel defect detection by using deep learning." In Journal of Physics: Conference Series, vol. 1755, no. 1, p. 012041. IOP Publishing, 2021.

3. Damacharla, Praveen, Achuth Rao, Jordan Ringenberg, and Ahmad Y. Javaid. "TLU-net: a deep learning approach for automatic steel surface defect detection." In 2021 International Conference on Applied Artificial Intelligence (ICAPAI), pp. 1-6. IEEE, 2021.
4. Boikov, Aleksei, Vladimir Payor, Roman Savelev, and Alexandr Kolesnikov. "Synthetic data generation for steel defect detection and classification using deep learning." Symmetry 13, no. 7 (2021): 1176.
5. Qian, Kun. "Automated detection of steel defects via machine learning based on real-time semantic segmentation." In Proceedings of the 3rd international conference on video and image processing, pp. 42-46. 2019.
6. Chigateri, Keerthana B., and Ajit M. Hebbale. "A steel surface defect detection model using machine learning." Materials Today: Proceedings (2023).
7. Akhyar, Fityanul, Ying Liu, Chao-Yung Hsu, Timothy K. Shih, and Chih-Yang Lin. "FDD: a deep learning–based steel defect detectors." The International Journal of Advanced Manufacturing Technology 126, no. 3 (2023): 1093-1107.
8. Teli, Shivagond & Majali, Vijay & Bhushi, Umesh. (2010). Role of Cost of Quality in the Automotive Industry.
9. Parnami, Archit & Lee, Minwoo. (2022). Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. 10.48550/arXiv.2203.04291.
10. Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-Shot Learning. ACM Computing Surveys.
11. Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J., & Isola, P. (2020). Rethinking Few-Shot Image Classification: A Good Embedding is All You Need? ECCV.
12. Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-Learning in Neural Networks: A Survey. IEEE TPAMI.
13. Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical Networks for Few-Shot Learning. Advances in Neural Information Processing Systems (NeurIPS), 30.
14. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning (ICML)..