The Categorization of Surface Irregularities Presents on the Hot-Rolled Steel Strip, Encompassing Six Distinct Types of Surface Anomalies, Achieved through the Implementation of Vision Transformer

Naimur Asif Borno and Anik Ghosh
Department of Mechatronics Engineering
Rajshahi University of Engineering & Technology (RUET)
Rajshahi, Bangladesh
naimurborno@gmail.com, anikghoshr123@gmail.com

Durjoy Datta Mazumder
Department of Materials Science & Engineering
Rajshahi University of Engineering & Technology (RUET)
Rajshahi, Bangladesh
durjoydmazumder1813058@gmail.com

Abstract

Amidst the epoch of the fourth industrial revolution, a discernible imperative surface within the steel industry necessitated the replacement of the archaic Defect Inspection System (DIS). The profound fiscal repercussions stemming from substandard steel underscore the exigency for this transition. Real-time diagnostics, a pivotal facet of quality control in manufacturing, grapples with inherent challenges, notably low automation and inconsistent flaw detection on steel surfaces. In response, a groundbreaking approach has materialized in the form of machine vision-based models, strategically devised to surpass the capabilities of conventional DIS and elevate the quality of produced steel. In the course of our study, we addressed flaws in six hot-rolled steel predicaments, leveraging a dataset encompassing ten critical surface defects: inclusion, pitted surface, crazing, rolled-in scale, patches, and scratches, thereby confronting the challenges previously articulated. Upon meticulous analysis of the dataset, our model, the Vision-Based Transformer (VIT), attained an exceptional accuracy rate of 98%. Four distinct machine learning models—Xception, ResNet50V2, EfficientNetB2, and MobileNetV2—were enlisted for performance evaluation, ultimately revealing the superiority of the VIT in the domain of vision-based Defect Inspection Systems.

Keywords
Steel Defect, Transformer, Vision Transformer, Computer Vision,ViT

1. Introduction

In terms of quantity and versatility of application, steel is arguably the most significant metal. The rise of industrial society has benefited greatly from steel. The most crucial quality factors, especially for items made of flat-rolled steel, are surface characteristics coupled with other attributes (Neogi et al 2014). These steels are employed in the production of various industrial machinery. The most crucial step in lowering the risk to safety and monetary loss is the identification of industrial machinery and associated workpieces. Surface imperfections differ amongst workpieces. There are various defects in the surface of steel plates such as "pitting", "burr", "scratch", "crack" etc. These flaws pose a risk to consumer safety in addition to lowering product quality. As a result, a popular area of study is the classification of surface defects in industrial workpieces. Workpiece defects can originate from a variety of industries, including manufacturing, construction, and electrical work (Li. et al 2022). Hot-rolled steel stands as a pivotal constituent in construction applications owing to its robustness and economic viability. Commonly available hot-rolled steel is produced by rolling steel at temperatures above 1700°F, which is far higher than the temperature at which most steels recrystallize. Steel products subjected to hot-rolling processes are known to have an extremely high rate of surface flaws of various kinds during the manufacturing process. Because of human error, manual strip inspection during production is labor-intensive and prone to mistakes. Therefore, automated techniques are preferred in order to reasonably confidently ensure the consistency of examination. In particular, Steel strip images can be acquired utilizing specific sensing hardware and subsequently analyzed employing specialized computer algorithms, employing image processing and computer vision methodologies. To guarantee a surface free of defects, traditional manual surface inspection techniques are woefully inadequate (Konovalenko et al 2020). The quality of steel bands is lowered by surface imperfections, but the way these damage types are classified allows for a prompt identification and removal of the underlying causes. Thus, the secret to metal product quality control is the effectiveness and precision of defect classification. Recently, a large number of optical-digital systems have been developed that enable defectoscopy of the surface of the rolled metal at a high enough level. A myriad of neural network architectures, including but not limited to GoogLeNet, AlexNet, and ResNet, are strategically harnessed to address a spectrum of challenges within the domain of defectoscopy. Its speed is determined by the model's complexity. neural systems are taught using pictures of specific metallurgical plant flaws (Ashour et al 2019). Driven by the remarkable achievements in natural language processing, computer vision transformers, which can prove to be highly advantageous in the domain of surface defect detection in metals.

Transformers are models that carry out a self-attention mechanism by giving each component of the incoming data a unique weight. Transformers are now the go-to models for jobs involving Natural Language Processing (NLP). Due to its scalability and computational efficiency, models can undergo training with an excess of 100 billion parameters without encountering performance saturation. Proposing an application of the same architecture, albeit with minor adjustments, for image classification is motivated by the proven effectiveness of transformers in natural language processing. The underlying assumption is that the self-attention mechanism, which has demonstrated utility in language tasks, could also prove advantageous in image classification endeavors. Transformers emerge as a favorable option for image-related tasks due to their capability to simulate long-range dependencies, adapt to diverse input sizes, and potentially process data in parallel (Maurício et al 2023). These days, vision transformers are widely used and have achieved great success in a variety of domains, including object identification, semantic segmentation, pose estimation, image and video classification, and posture estimation (Ruan et al. 2022). In contrast to CNN models that heavily depend on convolutional layers, the vision transformer utilizes the transformer architecture, a deep neural network grounded in the self-attention mechanism. Originally introduced in natural language processing (NLP), the transformer structure has been expanded into computer vision.

When compared to CNN models, those built on the transformer architecture demonstrate superior performance in image classification (Jiang et al. 2021). If we consider the merits of Vision Transformer (ViT) over alternative approaches in computer vision tasks, it becomes evident that ViT employs key strategies rendering it well-suited for ultimate classification challenges. Unlike conventional CNN models, where input size can pose a significant impediment, ViT adeptly addresses this issue through linear scalability. ViTs leverage self-attention mechanisms, enhancing their ability to effectively capture long-range dependencies in images. Furthermore, their capacity to process images as sequences of patches facilitates parallelization and efficient utilization of hardware resources. Notably, ViTs exhibit robust performance across diverse image tasks without reliance on handcrafted features. Additionally, their scalability allows seamless handling of inputs of varying sizes without necessitating extensive architectural modifications. These factors collectively establish ViT as exceptionally compatible with image vision tasks. In our research, we utilized a dataset pertaining to surface defects in hot-rolled steel strips. The accuracy of a model applied to such a dataset is paramount, making ViT a compelling choice for this computer vision task due to its demonstrated efficacy. Unlike traditional convolutional neural networks (CNNs), ViT's reliance on self-attention mechanisms for capturing global dependencies in image data represents an innovative paradigm that has yielded success in tasks such as object detection, image classification, and segmentation. ViT's versatility and scalability position it as an optimal solution for a myriad of real-world challenges across industries like healthcare, manufacturing, and autonomous systems. Consequently, ViT emerges as a promising and highly suitable solution for addressing the complexities inherent in industrial applications.

1.1 Objectives

This study aims to develop a vision transformer-based classification model for swiftly identifying surface defects in hot-rolled steel strips. The goal is to automate defect detection, reduce human intervention, and minimize operational costs. The approach involves formulating a classification model inspired by natural language processing, applying it to a curated dataset of steel strip anomalies, and rigorously assessing its accuracy against contemporary models. The anticipated outcome is an industry-compliant model that enhances defect detection while lowering operational expenses.

2. Literature Review

Autocorrelation is a statistical method used in defect detection systems (Cui et al., 2023) Other methods for defect detection include the random field model (Chen et al., 2019) hybrid and complementary fractal feature vector(Zhou et al., 2017) , and 2D-Wavelet Transform(Ghazvini et al., 2009) . Deep learning techniques, which have more benefits than standard approaches, provide an automated and proactive approach to surface flaw identification that may successfully replace traditional human inspection systems.The introduction and use of Deep Convolutional Neural Networks (DCNNs) to the task of defect classification in hot-rolled steel is a state-of-the-art development in this sector in recent years.In order to accomplish quick and accurate fault detection on steel surfaces,(Fu et al., 2019) have introduced a succinct but effective Convolutional Neural Network (CNN) model that combines numerous sensory fields and strongly emphasizes training of essential features at lower levels. It should be emphasized that the suggested technique may require a considerable amount of labeled data for training, a possible difficulty in certain cases where data gathering could be restricted. (Feng et al., 2021) adopted the RepVGG model to detect surface defects in hot-rolled steel. They have found an outstanding accuracy of 95.10% and use X-SSD datasets of 1350 images to evaluate model performance. but the RepVGG+SA algorithm has high computational complexity, which leads to high deployment costs. (Luo & He, 2016) concentrate on the architecture, implementation, and assessment of the system, which employs image processing and FPGA methods to identify and categorize faults on steel surfaces in real time with a 92.11% average accuracy. CNNs automate the acquisition of surface fault characteristics, boosting efficiency over human techniques.

Nevertheless, their particular sensitivity limits their capacity to understand the complete input data. To better overall feature characterization, bigger convolutional kernels, and more advanced layers are necessary, however, this could increase the level of complexity and possibly delay training convergence, resulting in a dimensional disaster. To address the issues stemming from the convolutional operations' inherent bias in CNNs, researchers have turned to the Transformer model in computer vision for image classification. The Transformer model, (Vaswani et al., 2017), has demonstrated its prowess by achieving top-tier results in machine translation tasks. Notably, it exhibits superior parallelizability and markedly reduced training time requirements. Consequently, the Transformer has emerged as a mainstream algorithm in the domain of natural language processing. Building upon this inspiration, the Vision Transformer (ViT) (Dosovitskiy et al., 2021) presents an innovative approach by directly applying the standard Transformer architecture to image patches, eliminating the need for convolutional layers. The paper demonstrates ViT's capacity to attain leading performance in image classification tasks. Notably, ViTs exhibit efficient training characteristics by demanding fewer parameters compared to CNNs, marking a significant advancement in the field of computer vision. In our study, we perform the categorization of steel surface flaws via a vision-based transformer. By training the Vision Transformer (ViT) model with a steel defect dataset, we can successfully recognize and classify flaws, assuring quality control in steel manufacturing. This demonstrates the promise of transformer-based models in image recognition applications. Our study stresses the vital significance of precise defect detection and object identification in preserving steel product quality

3. Methods

The detailed methodology is elucidated in Figure 1, which is divided into three main phases: preprocessing, model creation, and evaluation. Each of these phases is further segmented into specific sub-phases. The data acquisition phase has been discussed in Section 4, and the subsequent sections of this paper delve into a thorough examination of the remaining phases.

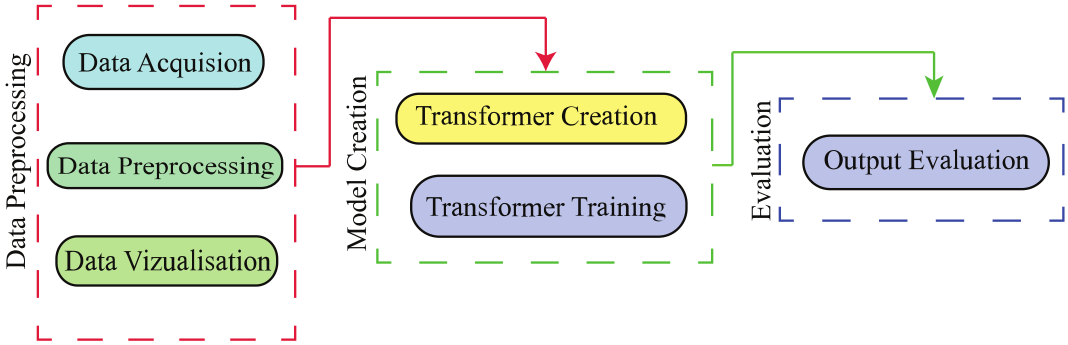

Figure 1. Flow diagram of the overall system.

3.1 Data Preprocessing

The comprehensive dataset encompasses 1728 monochromatic images intricately documenting discernible anomalies present on metallic surfaces. These images are systematically arranged into six carefully curated folders, namely Crazing, Inclusion, Rolled, Patches, Pitted, and Scratches. Each folder encapsulates a diverse array of images, each representing unique defect classes. Figure 2 provides a visual representation of the images contained in these different folders.

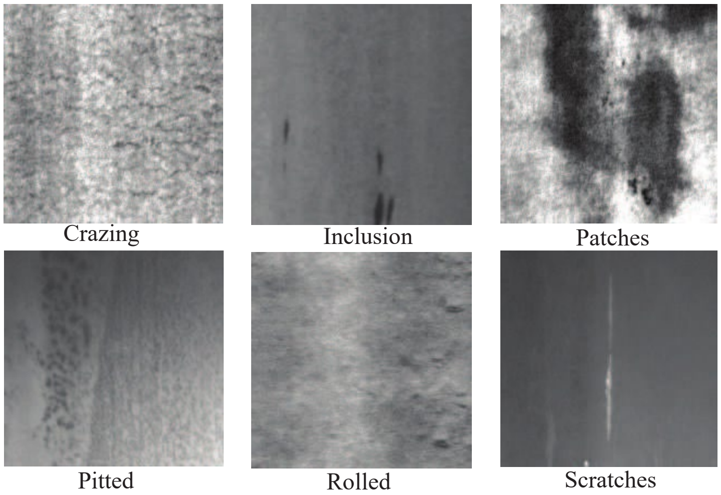

Figure 2. Hot Rolled Surface defects of different classes each

The 1728 images are standardized to dimensions of 360×360 pixels. Analysis of the histogram, generated from a representative subset of the dataset, reveals that pixel values range from 0 to 255. To achieve normalization, the array values of each image are divided by the maximum pixel value present within that particular image. Figure 3 provides a visual representation of the histogram distribution of the images.

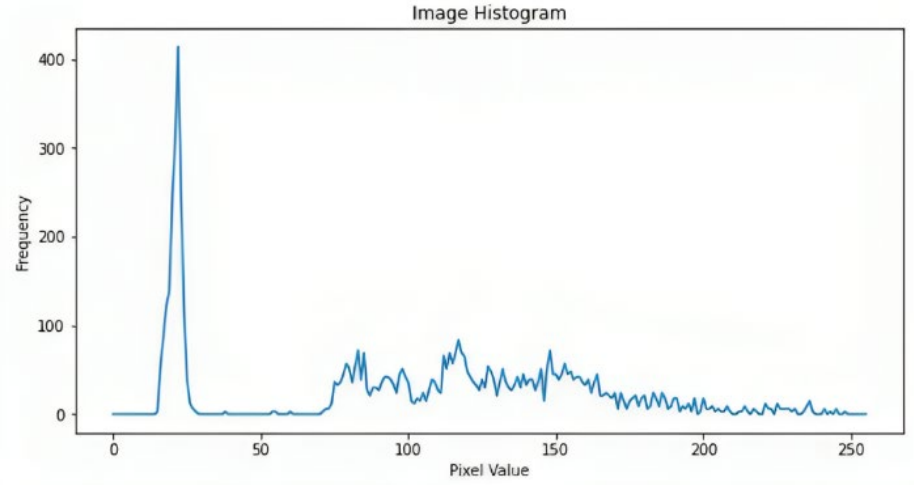

Figure 3. Histogram of Images

3.2 Vision Transformer

The Transformer framework, initially crafted for natural language processing, is characterized by its influential self-attention mechanism. Transitioning this paradigm to the realm of vision, the Vision Transformer's architectural intricacies are meticulously illustrated in Figure 4. The primary input to this transformative network comprises original grayscale images, each boasting dimensions of 360×360 pixels. However, before integration into the Transformer, a pivotal transformation takes place, involving the partitioning of images into multiple patches. These individual segments are then embedded to seamlessly integrate into the network. Following this patch-based embedding, a critical turn occurs with the introduction of a Flatten layer. This layer adeptly projects the flattened image segments, setting the stage for linear projection – a precursor to the subsequent transformation stage: positional embedding. In the realm of positional embedding, spatial information is infused into the patches, with dimensions meticulously determined by a well-defined equation (Hen et al., 2022):

$$pe(position, 2i) = sin\left(\frac{position}{10000^{\frac{2i}{dim}}}\right) \tag{1}$$

$$pe(position, 2i+1) = cos\left(\frac{position}{10000^{\frac{2i}{dim}}}\right) \tag{2}$$

Position, meticulously defined as a word's placement within a sentence, and 'i,' signifying the current dimension of positional encoding, play pivotal roles in this intricate framework. Following the indispensable steps of positional embedding and linear projection, the subsequent phase unfolds across eight Transformer encoder layers. Each layer initiates with the integration of a normalization layer,

$$\mu = \frac{1}{H}\sum_{i=1}^{H} a_i \tag{3}$$

$$\sigma = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i - \mu)^2} \tag{4}$$

where 'H' denotes the number of hidden units within each layer. The normalization layer, governed by equations (3) and (4), orchestrates a shared normalization for all hidden units in a layer. Notably, different training cases carry distinct normalization terms, adding a nuanced layer of complexity to the process. The normalized output progresses into the multi-head attention layer, symbolizing a symphonic crescendo in the integration of components. Each layer assumes a pivotal role in the intricate choreography of feature extraction and transformation. Following its traverse through this harmonious cascade of layers, the encoded output navigates through two layers of artificial neural dense architecture. The output bifurcates into two distinct sections: one dedicated to image classification. Given six distinct classes, the classification section comprises six dense layers, each employing the softmax activation function using the equation (5),

$$softmax(Z)_i = \frac{e^i}{\sum_{j=1}^{k} e^{z^j}} \tag{5}$$

Here the Z_i is the input to the softmax function for class i and k is the total number of classes. This architectural symphony is visually depicted in Figure 4.

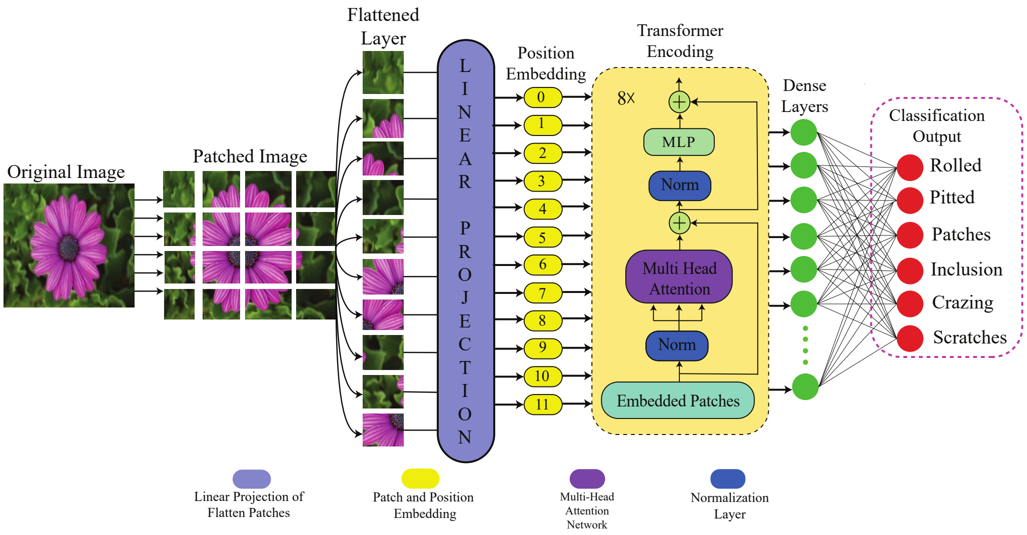

Figure 4. Architecture of Transformer Model

3.3 Evaluation Metrics

The model, composed of a sophisticated architecture featuring 8 layers of Transformer encoding, has undergone a rigorous training process within the Google Colab IDE. The training epochs meticulously honed the model's parameters, creating a refined configuration. In the realm of model evaluation, a comprehensive suite of metrics including accuracy, Mean Squared Error (MSE) loss, and Top-5-Accuracy has been deployed to assess its effectiveness. Accuracy, the primary metric among these, serves as a barometer for the model's precision in rendering predictions. It quantifies correctness by measuring the ratio of accurately classified instances to the total dataset size, expressed mathematically as in equation 5:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \qquad (5)$$

MSE provides additional insights into the variability of errors. The MSE formula is defined as in equation 6:

$$Mean\ Squared\ Error = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

where n is the number of observations, y_i is the actual value, $\hat{y}_i$ and is the predicted value. The metric denoted as "Top-5 accuracy" finds its prevalent application in the realm of multi-class classification challenges, specifically tailored for scenarios necessitating the identification of the most probable class amid the top quintile of predicted classes. This metric represents a nuanced expansion beyond the conventional top-1 accuracy, accommodating a broader spectrum of potential correct predictions.

$$Top - 5\ Accuracy = \frac{Number\ of\ test\ samles\ where\ True\ class\ is\ in\ Top\ 5\ prediction}{Total\ Number\ of\ Test\ Samples} \qquad (7)$$

The metric denoted as "Top-5 accuracy" finds its prevalent application in the realm of multi-class classification challenges, specifically tailored for scenarios necessitating the identification of the most probable class amid the top quintile of predicted classes. This metric represents a nuanced expansion beyond the conventional top-1 accuracy, accommodating a broader spectrum of potential correct predictions.

4. Data Collection

Derived from the NEU Metal Surface Defects repository, this dataset comprises a compilation of six distinctive surface irregularities observed in hot-rolled steel strips, specifically identified as rolled-in scale, patches, crazing, pitted surface, inclusion, and scratches. Comprising 1,800 grayscale images, the dataset is meticulously organized into three directories: one designated for training, another for testing, and the final one for validation. Within each directory, a consistent number of images representing each defect class is meticulously curated.

Table 1. Distribution of images from different classes

Class	Count (Training)	Count (Testing)	Total (Count)	Percentage
Crazing	276	12	288	16.67%
Inclusion	276	12	288	16.67%
Patches	276	12	288	16.67%
Pitted	276	12	288	16.67%
Rolled	276	12	288	16.67%
Scratches	276	12	288	16.67%

An analysis of the tabulated data reveals that the testing directory houses 12 image instances for each defect class. Conversely, the training directory boasts a more substantial compilation, featuring 276 image instances for each defect class. The validation directory, mirroring the testing directory, accommodates 12 image instances for each defect class. However, the model was constructed without considering the image data from the validation directory; this set of images was excluded from the building process. The image dimensions are a crucial aspect of data gathering. Each image has a size of approximately 360 x 360 x 3. In essence, it is imperative to elucidate that each defect class encompasses a total of 288 image instances, accounting for training and testing collectively. This distribution equates to 16.67% of the entire image dataset per defect class, demonstrating a thorough and systematic approach to the organization of the data across the various directories. Table 1 comprehensively presents the information pertaining to all the classes. The figure below clearly indicates that each defect class comprises an equal quantity of image data. It can be inferred from the figure that all defect classes consist of 288 image data points, and each class represents the same percentage of the overall image data.

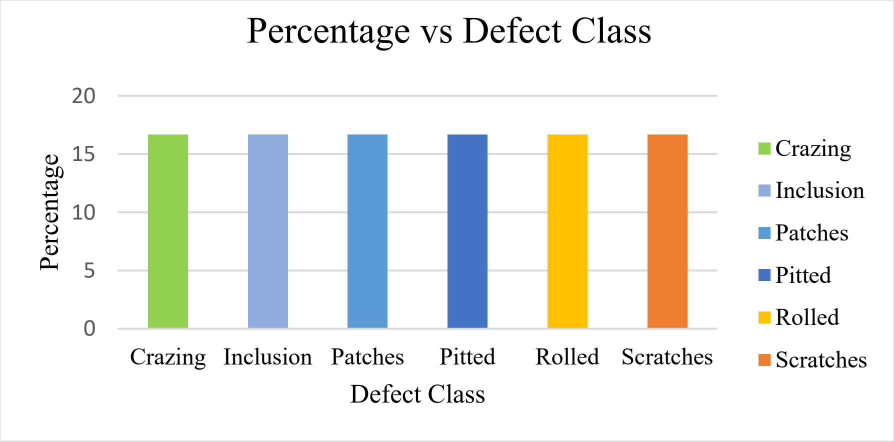

Figure 5. Distribution of images from different classes

5. Results and Discussion

5.1 Numerical Results

Table 2 presents a meticulous comparative analysis of diverse trained models, encompassing CNN with transfer learning through architectures like Xception, MobileNetV2, ResNet50V2, EfficientNetB2, and our Vision Transformer model. Within this tableau, the Xception model manifests the least favorable performance, achieving a classification accuracy of 91.3% alongside a classification loss of 0.521. In stark contrast, the EfficientNetB2 model emerges as the epitome of performance, boasting a noteworthy accuracy of 97.3% coupled with a classification loss of 0.210. MobileNetV2, with a moderate accuracy of 97%, exhibits a classification loss of 0.295. The crowning achievement within the table, however, is reserved for our Vision Transformer-based model, showcasing unparalleled results with an accuracy of 98% and a classification loss of 0.172. This comprehensive comparison underscores the superior efficacy of the Vision Transformer model in the context of classification accuracy and loss.

Table 2. Comparative analysis of Various models with our Model

Model	Classification Accuracy	Loss	Top-5-Accuracy
Our Model	**98%**	**0.172**	**99.8%**
CNN with Xception	91.3%	0.521	99%
CNN with ResNet50V2	94.6%	0.329	99.3%
CNN with MobileNetV2	97%	0.295	98%
CNN with EfficientNetB2	97.3%	0.210	99%

5.2 Graphical Results

Having undergone an extensive training regimen spanning 700 epochs, the Transformer model was subjected to dataset partitioning following the methodology explicated in Section 4. The quantitative outcomes, meticulously detailed in Table 2 of Section 5.1, unequivocally demonstrate that the model proposed in this paper outshines the performance of all referenced models in the aforementioned table. To further illuminate these findings, the graphical representation in Figure 6 delineates the comprehensive reduction in loss, comparing Training loss to Validation loss. Specifically, the ultimate convergence of training loss to 0.172 stands in stark contrast to the validation loss stabilizing at 0.276.

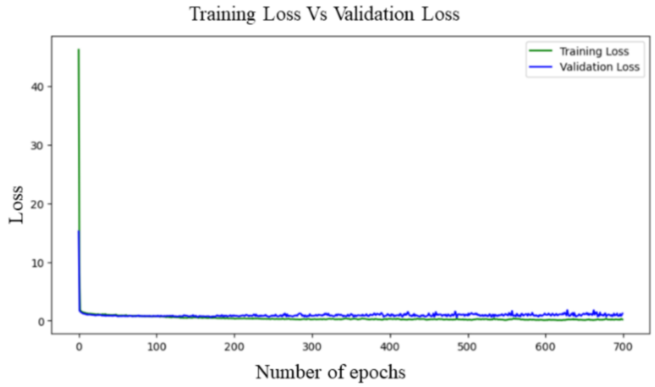

Figure 6. Training Loss vs Validation Loss

In tandem, Figure 7 provides a visual narrative of the training accuracy versus validation accuracy. The discerned trends unveil an impressive 99.8% training accuracy, juxtaposed with the validation accuracy plateauing at approximately 95.6%. The nuanced fluctuations in accuracy throughout the epochs are perceptibly evident in this graphical depiction.

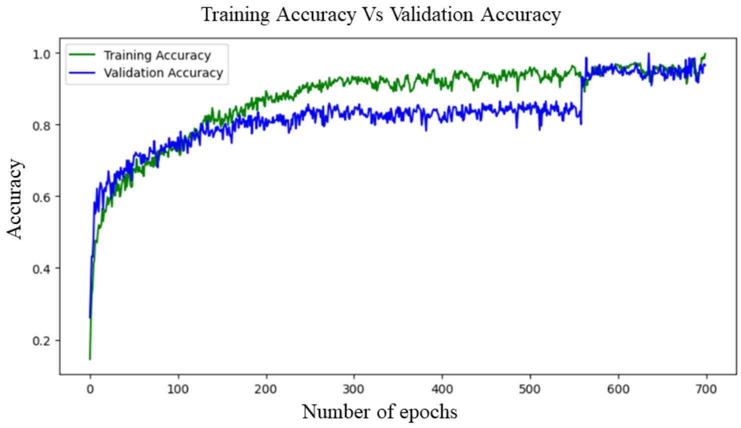

Figure 7. Training Accuracy vs Validation Accuracy

Furthermore, Figure 8 supplements these insights by illustrating the top-5-accuracy, offering a holistic perspective on the model's classification performance beyond the singular top-1 accuracy metric. These comprehensive analyses collectively underscore the robust performance and generalization capabilities of the proposed Transformer model.

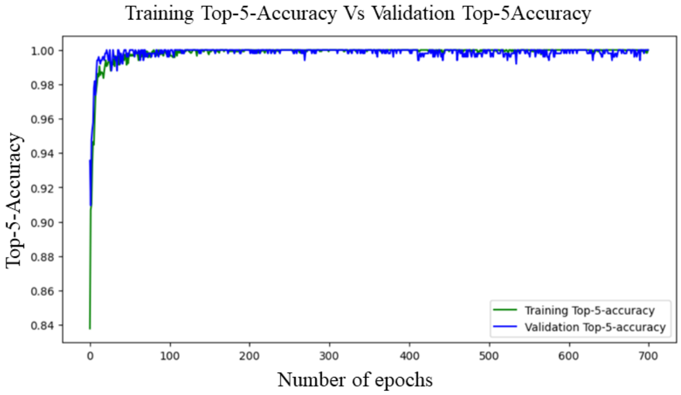

Figure 8: Training Top-5-Accuracy vs Validation Top-5-Accuracy

5.3 Validation

Figure 9 visually encapsulates the representation of predicted image classes, wherein the labels discerned are indicative of the contrast between predicted and actual labels. This graphical depiction serves as a visual testament to the model's accuracy in real-world image prediction, offering a nuanced insight into the model's efficacy in translating predictions to real-life scenarios.

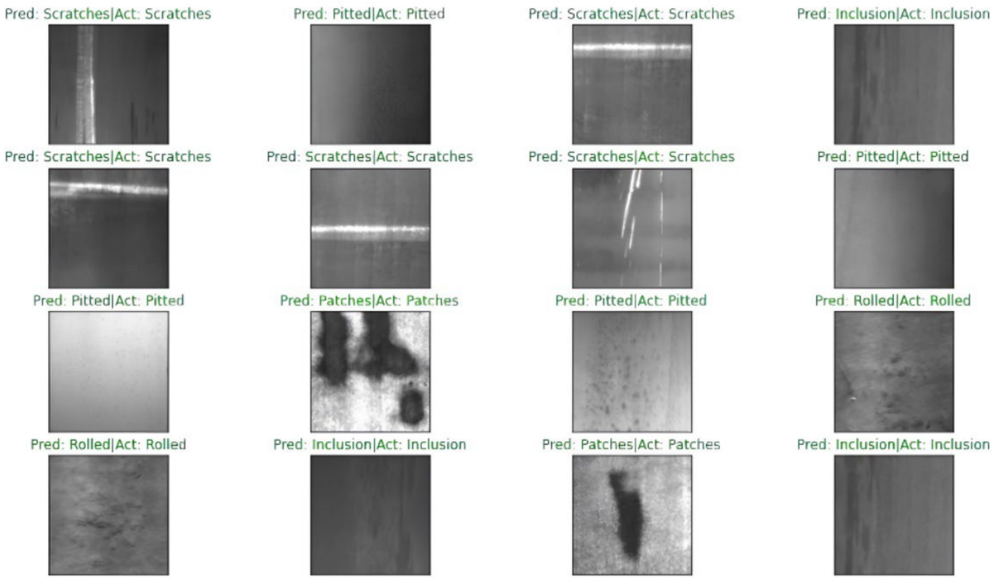
Pred: Scratches|Act: Scratches
Pred: Pitted|Act: Pitted
Pred: Scratches|Act: Scratches
Pred: Inclusion|Act: Inclusion
Pred: Scratches|Act: Scratches
Pred: Scratches|Act: Scratches
Pred: Scratches|Act: Scratches
Pred: Pitted|Act: Pitted
Pred: Pitted|Act: Pitted
Pred: Patches|Act: Patches
Pred: Pitted|Act: Pitted
Pred: Rolled|Act: Rolled
Pred: Rolled|Act: Rolled
Pred: Inclusion|Act: Inclusion
Pred: Patches|Act: Patches
Pred: Inclusion|Act: Inclusion

Figure 9. Prediction of Image classes

6. Conclusion

In the contemporary landscape, the domain of computer vision tasks has evolved into a complex realm, captivating both scholarly exploration and industrial applications. The demand for heightened accuracy in computer vision tasks has become imperative. In the course of our investigation, we have introduced a meticulously crafted neural network architecture tailored for the intricacies of computer vision tasks. The incorporation of the Vision Transformer into our model has proven instrumental, particularly in discerning and classifying surface defects. Through rigorous comparative analyses with alternative models, we sought to assess the model's adaptability to real-time challenges. The attained accuracy is notably commendable, boasting a training accuracy of 98% and an even more remarkable testing accuracy of 95%. These precision metrics surpass the requisites for adeptly tackling intricate computer vision tasks. Various deep learning models, encompassing CNNs with Xception, ResNet50V2, MobileNetV2, and EfficientNetB2, underwent rigorous evaluation alongside our proposed model. Ultimately, our model showcased superior adaptability, marked by its highest training and testing accuracy values. In summary, our model stands as a testament to its efficacy in executing classification tasks with unparalleled precision. Its applicability seamlessly extends to industrial contexts, proving invaluable in categorizing surface defects and proactively identifying their root causes, thereby facilitating timely remediation. This not only translates into economic savings by mitigating monetary losses but also serves as a judicious means of optimizing manpower costs. In essence, the deployment of the Vision Transformer model transcends mere efficiency, presenting itself as a boon to industries grappling with the temporal and financial challenges associated with manual detection processes.

References

Ashour, M. W., Khalid, F., Abdul Halin, A., Abdullah, L. N., & Darwish, S. H., Surface defects classification of hot-rolled steel strips using multi-directional shearlet features. *Arabian Journal for Science and Engineering*, 44, 2925-2932, 2019.

Chen, H., Liu, J., Wang, S., & Liu, K., Robust Dislocation Defects Region Segmentation for Polysilicon Wafer Image With Random Texture Background. *IEEE Access*, 7, 134318–134329, 2019.

Cui, W., Song, K., Feng, H., Jia, X., Liu, S., & Yan, Y., Autocorrelation Aware Aggregation Network for Salient Object Detection of Strip Steel Surface Defects. *IEEE Transactions on Instrumentation and Measurement,* 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arXiv:2010.11929), 2021.

Feng, X., Gao, X., & Luo, L., X-SDD: A New Benchmark for Hot Rolled Steel Strip Surface Defects Detection. *Symmetry*, 13(4), Article 4, 2021.

Fu, G., Sun, P., Zhu, W., Yang, J., Cao, Y., Yang, M. Y., & Cao, Y., A deep-learning-based approach for fast and robust steel surface defects classification. *Optics and Lasers in Engineering,* 121, 397–405, 2019.

Ghazvini, M., Monadjemi, S. A., Movahhedinia, N., & Jamshidi, K., Defect Detection of Tiles Using 2D-Wavelet Transform and Statistical Features.*World Academy of Science, Engineering and Technology* 49, 2009.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D., A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 87-110, 2022.

Jiang, Z., Dong, Z., Wang, L., & Jiang, W., Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model. *Computational Intelligence and Neuroscience*, 2021.

Konovalenko, I., Maruschak, P., Brezinová, J., Viňáš, J., & Brezina, J., Steel surface defect classification using deep residual neural network. *Metals*, *10*(6), 846, 2020.

Li, Q., Luo, Z., Chen, H., & Li, C., An overview of deeply optimized convolutional neural networks and research in surface defect classification of workpieces. *IEEE Access*, *10*, 26443-26462, 2022.

Luo, Q., & He, Y., A cost-effective and automatic surface defect inspection system for hot-rolled flat steel. *Robotics and Computer-Integrated Manufacturing,* 38, 16–30, 2016.

Maurício, J., Domingues, I., & Bernardino, J., Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, *13*(9), 5521, 2023.

Neogi, N., Mohanta, D. K., & Dutta, P. K., Review of vision-based steel surface inspection systems. *EURASIP Journal on Image and Video Processing*, *2014*(1), 1-19, 2014.

Ruan, B. K., Shuai, H. H., & Cheng, W. H., Vision transformers: state of the art and research challenges. *arXiv preprint arXiv:2207.03041,* 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I., Attention is all you need. *Advances in Neural Information Processing Systems,* 30, 2017.

Zhou, J., Wang, J., & Bu, H., Fabric defect detection using a hybrid and complementary fractal feature vector and FCM-based novelty detector. *Fibres & Textiles in Eastern Europe*, 6 (126), 46–52, 2017.

Biographies

Naimur Asif Borno is an undergraduate student pursuing studies in the Mechatronics Engineering department at Rajshahi University of Engineering & Technology. Demonstrating a keen interest in the domains of machine learning and computer vision, Borno has actively engaged in coursework and projects centered around these disciplines. His enthusiasm extends to the realms of robotics and control system work, reflecting a diverse array of technical interests. Notably, Borno is presently immersed in the research and development of his thesis, focusing on the intricate dynamics of microgrid systems. Additionally, he has augmented his academic pursuits with practical experience, having successfully completed an internship in machine learning at a Malaysia-based company, Malaysia. Complementing his theoretical knowledge with hands-on skills, Borno has accomplished computer vision projects, particularly in the area of object detection. This amalgamation of academic dedication, practical experience, and project work underscores Naimur Asif Borno's multifaceted commitment to advancing his expertise in the fields of mechatronics, machine learning, and computer vision.

Durjoy Datta Mazumder, an undergraduate student in the Materials Science & Engineering department at Rajshahi University of Engineering & Technology, stands out as a multifaceted researcher with a keen interest in diverse fields. His academic pursuits focus on machine learning, materials informatics, biomass, and biomaterials. Currently engrossed in his thesis on biomass and nanoparticles, Durjoy is exploring the intersection of these two domains. His commitment to interdisciplinary research is evident in his concurrent work on machine learning and computer vision tasks, specifically applied to material science challenges. Durjoy's hands-on experience extends to the industrial realm, having completed impactful training in a steel industry. This blend of academic rigor, research versatility, and practical industry exposure underscores his holistic approach to materials science and engineering.

Anik Ghosh is an ambitious undergraduate student enrolled in the Mechatronics Engineering department at Rajshahi University of Engineering & Technology. His fervent curiosity lies at the intersection of machine learning, control systems, and robotics. Currently immersed in his thesis, Anik is passionately exploring the realms of biomedical systems, showcasing a keen interest in the convergence of engineering and healthcare. Beyond academia, he has applied his skills to the field of aeronautical instruments, reflecting a versatile approach to engineering challenges. Anik boasts an outstanding academic record, a testament to his dedication and intellectual prowess. His endeavors extend to addressing machine learning-based problems within the medical sector, showcasing a commitment to leveraging technology for the betterment of healthcare practices. As he navigates the dynamic landscape of mechatronics, Anik Ghosh is poised to make meaningful contributions at the intersection of engineering innovation and healthcare advancement.