Comparative Analysis of CNN and Transformer models for image classification on Intel Image dataset.

Shreedatta Sawant
Assistant Professor, Department of Computer Engineering
Agnel Institute of Technology and Design
ssa@aitdgoa.edu.in

Ashish Narvekar
Assistant Professor, Department of Computer Engineering
Agnel Institute of Technology and Design
an@aitdgoa.edu.in

Aditya Pednekar
Student, Department of Computer Engineering
Agnel Institute of Technology and Design
adityapednekar1701@gmail.com

Dinesh Devalienaik
Student, Department of Computer Engineering
Agnel Institute of Technology and Design
naikd0153@gmail.com

Dhruv Malvankar
Student, Department of Computer Engineering
Agnel Institute of Technology and Design
malvankardhruv4@gmail.com

Abstract— This comparative study investigates the use of six neural network architectures-ResNet50, VGG16, EfficientNetB0, Vision Transformer, Swin Transformer and DenseNet -for classifying images in the Intel Image dataset. The research conducted tends to discover the disadvantages as well as advantages of each model in terms of accuracy, computational efficiency, and versatility. By implementing and fine-tuning these architectures on the chosen dataset, we assess their ability to categorize various image types. This study provides insights of the balance in between model complexity and as well as performance, offering valuable guidance for researchers and professionals in selecting appropriate neural network architectures for diverse image classification tasks.

Keywords : Deep learning, CNN, ViT , ResNet50, EfficientNet, VGG16.

I. Introduction

Image classification has become an important task in computer vision, with wide-ranging applications in fields such as healthcare, autonomous vehicles, and security systems[1][2][3]. As visual data becomes increasingly complex and voluminous, the need for efficient and accurate image classification models has grown significantly. Convolutional Neural Networks (CNNs) over the time have emerged as the leading approach for tackling image classification challenges, demonstrating remarkable results on various benchmark datasets[4][5].

This study examines three prominent CNN architectures and one transformer model: ResNet50, VGG16, EfficientNet, and Vision Transformer (ViT). These models represent significant landmarks for the evolution of DL(deep learning) to take place for computer vision tasks:

1. ResNet50: Introduced by He et al. (2016), ResNet (Residual Network) helped people to focus and address the issue of vanishing gradients which occurs in deep networks when they make use of skip connections[6]. The 50-layer ResNet50 architecture has been able to show exceptional performance in the task of image classification.

2.VGG16: It was created by Simonyan and Zisserman(2014), VGG16 is known for its simple design and depth[7]. It makes use of convolutional filters(3 X 3) which are smaller and stacked in increasing depth, emphasizing the importance of network depth in achieving high performance.

3. EfficientNet, created by Tan and Le (2019), introduced a novel method(scaling method) that uniformly scales networks based on width, depth, and resolution[8].State-of-the-art accuracy was achieved through this approach on various benchmarks while maintaining efficiency.

4. Vision Transformer (ViT) created by Dosovitskiy et al. (2020), applying a transformer-based architecture to image classification. ViT helps us in dividing the images in patches and the use of self-attention helps to model long-range relationships.

5.Swin Transformer is a hierarchical (Vit) vision transformer that uses self-attention within shifted windows. This method substantially lowers computational cost while preserving good representational capacity. Its design enables scalable modeling of high-resolution images and has been shown to be effective on many vision tasks by extracting local and global information.

6.DenseNet (Dense Convolutional Network) proposes dense connectivity in which every layer takes inputs from all the previous layers. This architecture promotes feature reuse, enhances gradient flow, and decreases the number of parameters. By directly connecting layers, DenseNet avoids the vanishing gradient issue and allows for more efficient and accurate training of deep neural networks.

The main goal of the study performed is to evaluate these four architectures using the Intel Image dataset, examining their effectiveness, efficiency, and adaptability. By implementing and optimizing each model on the selected dataset, we aim to provide a comprehensive analysis of their strengths and limitations.

ISBN: 978-93-343-1044-3

This research is motivated by the need to understand how different deep learning models perform on specific datasets, as model selection can significantly impact the outcome of image classification projects. The results of this study will help to provide advantages to researchers in computer vision, while also aiding in the selection of required models for various image classification tasks.

II. Literature Survey

ResNet50, a deep residual neural network, has demonstrated impressive results in tasks involving image classification [9], [10], [11]. Making use of residual connections helps address the issue of vanishing gradients and also provide help in training of more deeper models as compared to earlier CNNs [12]. This increased depth enables the extraction of complex features, enhancing accuracy on large and varied datasets [9]. However, this also leads to higher computational requirements for both training and inference [12]. The Research conducted points out that data augmentation can greatly enhance its accuracy and robustness [13], [14]. The performance can be depending on the dataset and task, and in some instances, other architectures may outperform it in terms of speed or accuracy [15].

VGG-16, another popular CNN, is characterized by a simple design with sequential convolutional and max-pooling layers [9], [10], [11]. Although it is not as deep as ResNet50, it performs effectively in image classification tasks [6] and offers quicker training and inference [3].The ability to learn complex features might be limited due to Its shallower architecture might limit its ability to learn complex features, potentially resulting in lower accuracy on challenging datasets [9]. Like ResNet50, its performance is influenced by data augmentation [10], [14] and the characteristics of the dataset [15]. The decision between the two often involves balancing VGG-16's speed against ResNet50's superior accuracy [15].

EfficientNetB0, part of a CNN family optimized for providing higher accuracy with significantly lesser parameters than ResNet50 and VGG-16 while the depth, width, and resolution being scaled up,[15],[16]. Its efficiency results in faster training, reduced memory usage, and suitability for real-time or resource-constrained environments [17]. Studies suggest it can particularly on large datasets surpass ResNet50 and VGG-16 in terms of both accuracy and as well as the efficiency,[15]. Nonetheless, its performance still relies on dataset characteristics, augmentation, and tuning [10], [14], and it may not excel in all tasks [15]

This review compared ResNet50, VGG-16, EfficientNetB0, and ViTs for outdoor object classification. CNNs like ResNet50 and VGG-16 provide strong accuracy and robustness, while EfficientNetB0 is efficient for resource-limited settings [15]. ViTs are adept at capturing long-range dependencies but face challenges with computational demands and variability in outdoor scenes [19]. The choice of architecture depends on application needs, resources, and datasets. Future research should focus on standardized benchmarks, hybrid models, and improving ViT efficiency. Broader comparisons across diverse datasets and object classes will help clarify each model's strengths and limitations, supporting progress in this evolving field.

Vision Transformers (ViTs) are increasingly utilized across industrial, medical, and agricultural sectors for their strong feature extraction capabilities. While CNNs have been leading in industrial inspection, ViTs now match or exceed them in complex, data-scarce tasks, such as railway freight car damage assessment [24]. In medicine, ViTs have outperformed CNNs in detecting diabetic retinopathy and segmenting brain tumors due to their non-local receptive fields [25], [26]. In agriculture, ViTs like ConvViT and Swin Transformer have demonstrated high accuracy in crop disease and rice classification [27], [28]. For security and surveillance, ViTs have enhanced human activity recognition and drone threat detection, achieving superior accuracy and efficiency over CNNs [29], [30]. These advancements underscore ViTs' growing applicability in real-world scenarios.

In recent times, there has been a growing interest in utilizing deep learning, especially convolutional neural networks (CNNs) and vision transformers (ViTs), for tasks related to medical image analysis. A significant innovation in this area is the creation of Swin Unet3D, which combines Vision Transformer elements with 3D convolutional frameworks to manage intricate volumetric data like brain MRIs [33]. This hybrid design leverages the spatial capabilities of convolutions and the comprehensive receptive field of transformers, leading to enhanced segmentation outcomes on the BraTS2021 dataset. Similarly, researchers have investigated improved EfficientNet architectures for detecting multi-grade brain tumors, achieving up to 98.6% accuracy while considerably lowering the number of parameters, underscoring EfficientNet's capacity to balance performance with computational efficiency [34]. A pivotal advancement in this field is the Swin Transformer, a hierarchical vision transformer that uses an innovative shifted window approach to make use of local attention linearly along with computational complexity which concerns image size [37]. This architecture allows it to effectively scale to high-resolution images and has shown strong results on various benchmarks, including ImageNet, COCO, and ADE20K. Building on this foundation, the introduction of Swin Transformer V2 represents a significant advancement in scalable vision modeling [35]. By utilizing techniques such as residual-post-norm with cosine attention and log-spaced position bias, Swin V2 enables effective training on ultra-high-resolution images, outperforming previous models on benchmarks like ImageNet and COCO.

Meanwhile, hybrid architectures that combine ViT and CNN components have gained popularity. For example, a model that integrates both approaches was proposed for detecting tuberculosis in chest X-rays, achieving a high classification accuracy of 91.1% and demonstrating the synergy between local (CNN) and global (ViT) feature extraction [36]. Additionally, studies have highlighted the importance of model optimization and data strategies. For instance, efforts to optimize the EfficientNet model not only focused on

ISBN: 978-93-343-1044-3

architecture but also included interpretability tools like Grad-CAM, enhancing clinical usability through transparent decision-making processes [34]. Similarly, employing data augmentation and focal loss in training hybrid models for chest X-ray anomaly detection has been shown to improve model generalization, particularly when dealing with imbalanced datasets [36]. These advancements collectively indicate a trend in model design where transformers, convolutions, and interpretability tools are increasingly being integrated to meet the demands of critical medical imaging tasks. While models like ResNet50 and EfficientNet remain reliable baselines, recent transformer-based and hybrid approaches—especially those based on Swin architectures—demonstrate promising improvements in both accuracy and efficiency, with an increasing emphasis on scalability and real-world clinical application.

III. Experiment

The study which is being conducted examines the comparative effectiveness of four leading deep learning models – ResNet50, VGG-16, EfficientNetB0, and Vision Transformers (ViTs) – in classifying outdoor object images. The research utilized a dataset approximately having a size of 25,000 images from From Intel image classification dataset(kaggle), categorizing them into six classes: Glacier, Mountain,Buildings, Sea,Forest, and Street.

The evaluation employed five key metrics: Confidence, Precision, Accuracy, Recall, and F1-score.

Confidence: This metric represents the model's level of assurance in its prediction, expressed as a probability.

Precision: It is a metric which is used to measure the correctness of favourable predictions made by a model. It is said to be the ratio of correct favourable predictions relative to the whole number of favourable predictions.

Accuracy: Among the simplest classification metrics, accuracy is computed by taking the correct predictions(no of predictions) and dividing it by the total predictions made.

Recall : It is utilised for assessing the capability of a model to faultlessly identify the suitable instances of a favourable class. The definition is as follows where the proportion of correct favourable predictions to the actual positive instances(Total number).

F-1 Score: It is a measure that unites both precision as well as recall into a single value and thus provides balanced results of a specific model's performance. It is identified as the harmonic average between 2 factors that are precision and recall.

TABLE I DETECTION METRICS FOR TRAINING OF MODELS

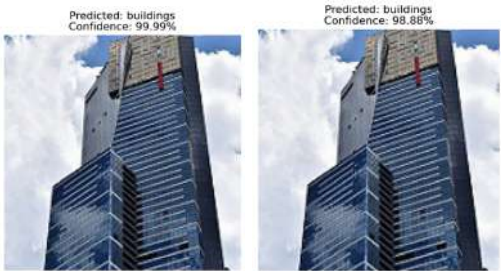
Methods	Accuracy	Precision	Recall	F1-Score
ResNet50	0.9143	0.9153	0.9170	0.9160
VGG16	0.8900	0.829	0.8920	0.8921
EfficientNet B0	0.8677	0.8701	0.8719	0.8703
ViT	0.9283	0.9299	0.9303	0.9288
Densenet	0.8827	0.8858	0.8853	0.8843
Swin	0.9257	0.9275	0.9273	0.9274

Table I displays the performance metrics for the six models evaluated. The (Vit) Vision Transformer performed well in all metrics, securing the top scores in Accuracy (0.9283), Precision (0.9299), Recall (0.9303), and F1-Score (0.9288). This outstanding performance underscores ViT's effective utilization of self-attention mechanisms to capture the global context so as to perform tasks in image classification . The Swin Transformer was a close second, delivering strong results with an Accuracy of 0.9257, Precision (Macro) of 0.9275, Recall (Macro) of 0.9273, and F1-Score (Macro) of 0.9274. It also achieved a high average confidence score of 0.9455, indicating reliable predictions. ResNet50 demonstrated competitive performance with an Accuracy of 0.9143, Precision of 0.9153, Recall of 0.9170, and F1-Score of 0.9160, highlighting the benefits of residual connections in mitigating vanishing gradients during deep network training. VGG16's results were moderately lower compared to ResNet50 and ViT, with an Accuracy of 0.8900, Precision of 0.8929, Recall of 0.8920, and F1-Score of 0.8921. DenseNet also showed solid performance, achieving an Accuracy of 0.8827, Precision (Macro) of 0.8858, Recall (Macro) of 0.8853, and F1-Score (Macro) of 0.8843, with an average confidence score of 0.8804, indicating fairly confident predictions across classes. Lastly, EfficientNet B0 recorded the lowest performance among the models, with an Accuracy of 0.8677, Precision of 0.8701, Recall of 0.8719, and F1-Score of 0.8703. Despite its lightweight design, EfficientNet B0 may need deeper variants to match the performance of more complex architectures in this task.

(a) ResNet50 (b) VGG-16

(c) EfficientNet B0 (d) ViT

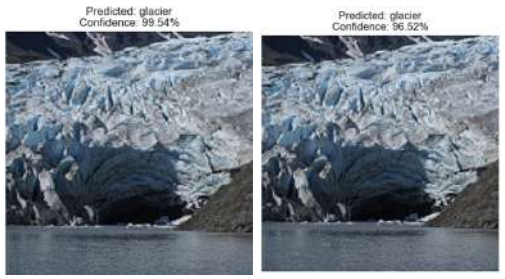
ISBN: 978-93-343-1044-3



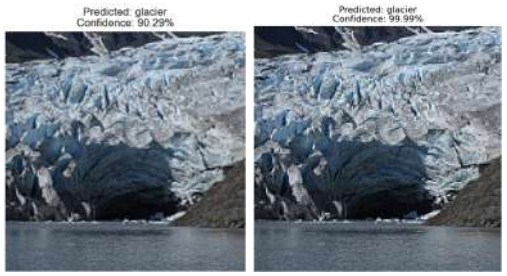
(e)Swin (f)Densenet

Fig. 1. Image Classification results for image 1(buildings class).

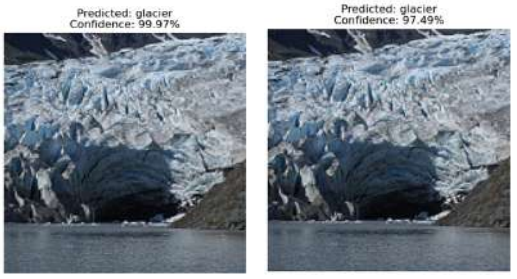
In a distinct analysis of an image from the "buildings" category (Fig. 1), VGG16 achieved the highest classification confidence at 99.99%. This was closely matched by both the Vision Transformer (ViT) and ResNet50, each scoring 99.98%, underscoring their reliability for this type of image. The Swin Transformer also performed exceptionally well, with a confidence score of 99.99%, equaling VGG16 and demonstrating its capability in capturing hierarchical features through shifted windows. DenseNet followed with a strong confidence level of 98.88%, highlighting its proficiency in feature reuse and dense connectivity. In contrast, EfficientNet-B0 showed a lower confidence of 94.97%, indicating that while it is efficient in terms of parameters and speed, it may not perform as well in high-detail scenarios like complex building structures. These confidence levels illustrate the varying strengths of different architectures in handling specific visual categories and emphasize the importance of choosing the right model based on the data's nature.



(a) Resnet50 (b) VGG-16



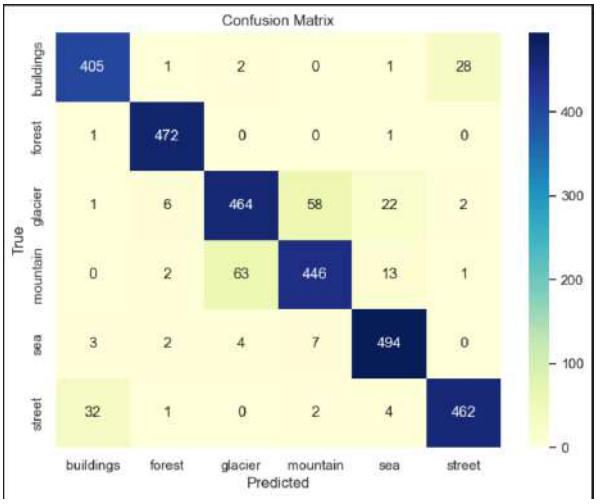
(c) EfficientNet B0 (d) ViT



(e)Swin (f)Densenet

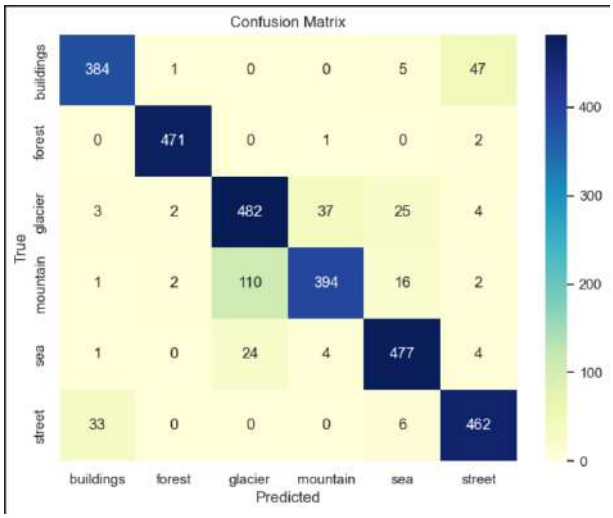
Fig. 2. Image Classification results of image 2(glacier class).

In a concurrent evaluation of the "glacier" class image, shown in Fig. 2, ViT achieved the highest classification confidence at 99.99%, demonstrating its remarkable accuracy for this class. The Swin Transformer followed closely with a confidence of 99.97%, showcasing its ability to capture fine-grained patterns through hierarchical attention. ResNet50 maintained strong performance with 99.54%, while DenseNet achieved a commendable 97.49%, utilizing its densely connected architecture for precise feature representation. VGG16 reached a classification confidence of 96.52%, indicating reliable results, and EfficientNet-B0 attained a confidence of 90.29%, reflecting slightly lower accuracy compared to the other models but still satisfactory for this image. These findings highlight the strengths of transformer-based models, particularly ViT and Swin, in managing intricate visual textures like glaciers.

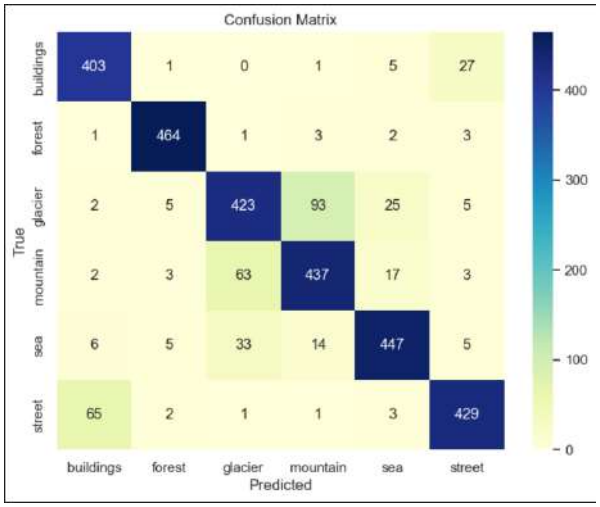


fig(3-a)ResNet50(confusion matrix)

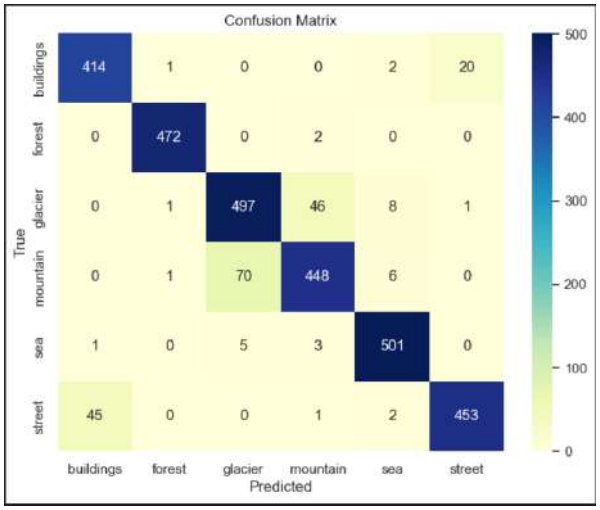
ISBN: 978-93-343-1044-3



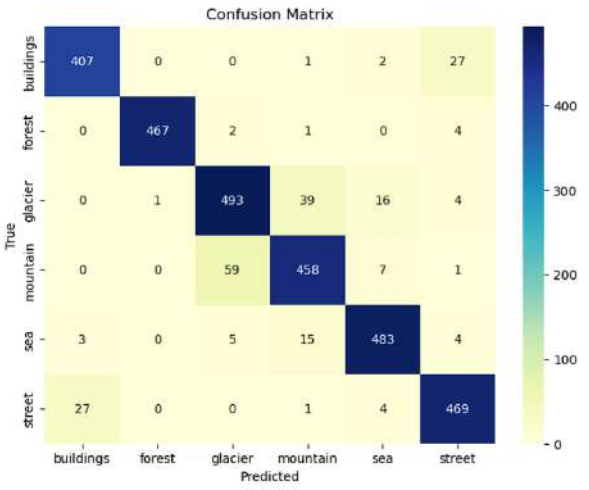
fig(3-b)VGG16(confusion matrix)



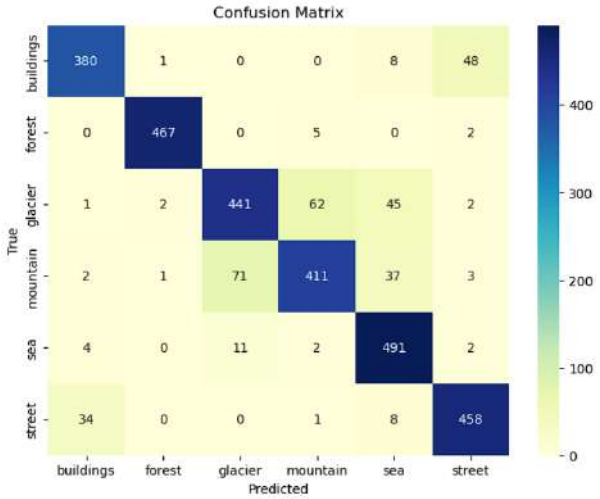
fig(3-c)EfficientNet B0(confusion matrix)



fig(3-d)ViT(Confusion Matrix)



fig(3-e)Swin Transformer(Confusion Matrix)



fig(3-f)DenseNet(Confusion Matrix)

IV. RESULTS

The performance indicators of six various image classification models: Vision Transformer (ViT), Swin Transformer, ResNet50, VGG16, DenseNet, and EfficientNet B0. Among them, ViT recorded the best performance on all the most important metrics, such as Accuracy (0.9283), Precision (0.9299), Recall (0.9303), and F1-Score (0.9288). This indicates the ability of ViT to well learn global dependencies in images with self-attention mechanisms.The Swin Transformer also exhibited good performance, second only to ViT with Accuracy of 0.9257, Precision (Macro) of 0.9275, Recall (Macro) of 0.9273, and F1-Score (Macro) of 0.9274. It also attained the highest mean confidence score (0.9455), suggesting confident and reliable predictions.

ResNet50 kept competitive performance levels with Accuracy (0.9143), Precision (0.9153), Recall (0.9170), and F1-Score (0.9160), confirming the strength of residual learning in deep convolutional networks. VGG16 and DenseNet performed moderately, with VGG16 performing marginally better than DenseNet across all metrics. VGG16 recorded an Accuracy of 0.8900, while DenseNet recorded

ISBN: 978-93-343-1044-3

0.8827. Their F1-Scores were 0.8921 and 0.8843, respectively.At the lower end of the scale, EfficientNet B0 had the worst results with an Accuracy of 0.8677 and F1-Score of 0.8703. Although it has a lightweight model, its performance indicates that deeper models might be required for more complicated classification problems.

Based on our literature review and experimental findings, we could conclude that ViT is particularly well-adapted for image classification tasks requiring high accuracy, especially when computational resources are not constrained. Secondly, the Swin Transformer ranked second with highly accurate (0.9257) and the best average confidence (0.9455), reflecting solid predictions. ResNet50 also performed highly, leveraging its residual learning architecture.VGG16 and DenseNet yielded average performance, with DenseNet falling behind VGG16 marginally. EfficientNet B0, being computationally effective, gave the poorest performance, which implies that more intricate tasks can demand deeper or stronger variants.

Deep learning improvements have dramatically enhanced image classification by allowing auto feature extraction and decision-making. Comparative research based on various architectures indicates that ResNet50 beats VGG16 and VGG19 in classifying products, with the maximum accuracy of 97.33% at epoch 20 [30]. Likewise, comparison of EfficientNet and MobileNetV2 with the Intel Image Dataset showed optimization methods like scaling techniques and automated mixed precision training enhanced EfficientNet to 94.5% accuracy and MobileNetV2 to 92% [31]. Another experiment analyzing the effect of learning rate adjustment reported that MobileNetV2 surpassed EfficientNet, reaching a high 99.67% accuracy at epoch 50, which confirms its effectiveness in sorting images. The results highlight the need for proper model choice, fine-tuning, and optimization methods to increase the accuracy of classification and computation speed for real-world tasks like product classification, object detection, and pattern identification [32].

V. Discussion

We developed and tested two different custom Convolutional Neural Network (CNN) architectures—named CustomCNN and DeepCNN —to determine how their performance would compare on the Intel Image Classification dataset. The main architectural variation is the depth and richness of the networks: the original CustomCNN has three convolutional blocks with a maximum of 128 filters, whereas the advanced DeepCNN architecture stretches to five convolutional blocks, with a maximum of 512 filters. Both models use ReLU activations and batch normalization following every convolutional layer. DeepCNN, however, has a stronger regularization approach with greater dropout (0.6) and more intense data augmentations in the form of random affine transformations, vertical flips, and color jitter, along with L2 regularization on the Adam optimizer. With regard to performance, CustomCNN gained a highest training accuracy of 72.43% and its highest validation accuracy of 83.87% with the lowest validation loss of 0.5148. However, DeepCNN obtained a slightly better training accuracy of 72.95% and attained a validation accuracy of 81.53%, along with the lowest validation loss of 0.4855. While CustomCNN reached a little higher validation accuracy, DeepCNN showed more stable and regular training behavior in later epochs, suggesting better generalization ability. Importantly, DeepCNN's validation loss fell more progressively, demonstrating the advantages of deeper feature extraction as well as greater regularization. Such results warrant the addition of both models for comparative investigation, showing the impact of architectural depth and regularization on model performance and aiding more complete assessment of CNN-based image classification methods.

Ensuring fairness in AI, particularly in image classification, is a significant issue, as biased training data can result in unfair outcomes, especially in critical areas like healthcare and security. Both CNNs and ViTs are susceptible to class imbalance, where classes with fewer examples experience reduced accuracy. To combat this, strategies such as adversarial debiasing, fairness-aware training, and data augmentation are utilized. Another ethical challenge is model explainability, as black-box models hinder transparency. Techniques in Explainable AI (XAI), fairness-oriented loss functions, and the curation of diverse datasets are crucial for the ethical deployment of AI.

Implementing AI in practical scenarios also presents challenges, including data privacy, vulnerability to adversarial attacks, and domain shift issues. Following the regulations like GDPR and HIPAA is crucial when managing sensitive image data. Research is ongoing to enhance robustness against adversarial perturbations and to improve generalization across domains. Additionally, scalability becomes a challenge when deploying CNNs and ViTs, particularly for high-resolution data. ViTs, in particular, demand large datasets, which limits their application in environments with limited data. Techniques such as model compression (pruning, quantization, distillation) and edge AI deployment help minimize computational demands, while hybrid CNN-ViT models provide a balance between accuracy and efficiency.

On the Intel Image Classification dataset, transfer learning was highly effective across all models by utilizing pre-trained ImageNet weights. Initial feature extraction results were strong, especially for ResNet50 and EfficientNet-B0. Performance improved further by selectively fine-tuning deeper layers, enabling models to better adapt to domain-specific features. Additionally, hyperparameter tuning—including learning rates, optimizers (Adam, AdamW), and schedulers—was crucial for convergence. Data augmentation and regularization methods like dropout and L2 weight decay also enhanced generalization. Overall, the combination of transfer learning, targeted fine-tuning, and optimized training strategies led to robust and scalable performance across both CNNs and ViTs.

VI. Conclusion

From all the research conducted and experiments carried out on the image classification models—ResNet50, VGG16, EfficientNetB0, ViT, Swin Transformer and DenseNet—the

ISBN: 978-93-343-1044-3

most accurate and consistent model for the work to be performed was identified to be the Vision Transformer. Its remarkable results, which were based on Accuracy, precision, recall, and F1-score metrics with a high prediction confidence for a particular class proved the validity of the transformation used in dealing with complex data.The Swin Transformer was also a strong performer, especially in confidence of prediction.ResNet50 also performed outstandingly, as it is the best alternative if a balance of computational efficiency with accuracy is needed. EfficientNetB0, being on the other side, was efficient with parameters but still had low confidence levels and relatively low accuracy compared to the rest, making it less competitive for highly accurate classifications. The classic model, VGG16, had achieved reasonable performance but lagged behind newer architectures.

Although conventional convolutional architectures such as ResNet50, VGG16, and DenseNet performed well, they were marginally surpassed by the transformer-based models. EfficientNet B0, although efficient, was behind when it came to overall performance. Deep or more complex variants of EfficientNet or hybrid models combining convolutional and transformer-based strategies could be explored in future work for better classification results.

In future work, training models like ViT , Swin Transformer ResNet50 on larger and more diverse datasets will help in further improving the classification performance and robustness. Accuracy and generalizability can be enhanced by using data augmentation techniques and fine-tuning pre-trained models on domain-specific datasets. This will enable these models for deployment in real-world applications, such as environmental monitoring, medical imaging, and autonomous systems, where accurate image classification is important..

References

1] C. Raptis, E. Karavasilis, G. Anastasopoulos, and A. Adamopoulos, “Comparative analysis of conventional CNN vs. ImageNet pretrained ResNet in medical image classification,” *Information*, vol. 15, no. 12, p. 806, 2024. doi: 10.3390/info15120806.

[2] [2] S. Rajaraman, S. K. Antani, M. Poostchi, K. Silamut, M. A. Hossain, R. J. Maude, S. Jaeger, and G. R. Thoma, “Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images,” *PeerJ*, vol. 6, p. e4568, 2018. doi: 10.7717/peerj.4568.

[3] D. Gyawali, A. Regmi, A. Shakya, A. Gautam, and S. Shrestha, “Comparative analysis of multiple deep CNN models for waste classification,” *arXiv preprint arXiv:2004.02168*, 2020. doi: 10.48550/arXiv.2004.02168.

[4] E. da Silva Puls, M. V. Todescato, and J. L. Carbonera, “An evaluation of pre-trained models for feature extraction in image classification,” *arXiv preprint arXiv:2310.02037*, 2023. doi: 10.48550/arXiv.2310.02037.

[5] P. Doungpaisan and P. Khunarsa, “A comparative study of pre-trained models for image feature extraction in weather image classification using Orange data mining,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 1, pp. 241–249, 2024. doi: 10.11591/ijeecs.v37.i1.pp241-249.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015. doi: 10.48550/arXiv.1512.03385.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. doi: 10.48550/arXiv.1409.1556.

[8] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *arXiv preprint arXiv:1905.11946*, 2019. doi: 10.48550/arXiv.1905.11946.

[9] N. Sharma, V. Jain, and A. Mishra, “An analysis of convolutional neural networks for image classification,” *Procedia Computer Science*, vol. 132, pp. 1341–1347, 2018. doi: 10.1016/j.procs.2018.05.198.

[10] K. Purnachand, S. Alghamdi, N. Veeraiah, Y. Alotaibi, S. Thotakura, and A. Alsufyani, “A new method for scene classification from the remote sensing images,” *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1–10, 2022. doi: 10.32604/cmc.2022.025118.

[11] M. Buyukdemircioglu, R. Can, and S. Kocaman, “Deep learning based roof type classification using very high resolution aerial imagery,” *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2021, pp. 55–61, 2021. doi: 10.5194/isprs-archives-xliii-b3-2021-55-2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015. doi: 10.48550/arXiv.1512.03385.

[13] L. Nanni, M. Paci, S. Brahnam, and A. Lumini, “Feature transforms for image data augmentation,” *Neural Computing and Applications*, vol. 34, no. 11, pp. 8971–8984, 2022. doi: 10.1007/s00521-022-07645-z.

[14] L. Nanni, M. Paci, S. Brahnam, and A. Lumini, “Comparison of different image data augmentation approaches,” *Journal of Imaging*, vol. 7, no. 12, p. 254, 2021. doi: 10.3390/jimaging7120254.

[15] B. Laschowski, W. McNally, A. Wong, and J. McPhee, “Environment classification for robotic leg prostheses and exoskeletons using deep convolutional neural networks,”

ISBN: 978-93-343-1044-3

Frontiers in Neurorobotics, vol. 15, p. 730965, 2022. doi: 10.3389/fnbot.2021.730965.

[16] A. A. Adegun, S. Viriri, and J. R. Tapamo, "Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis," *Journal of Big Data*, vol. 10, no. 1, p. 47, 2023. doi: 10.1186/s40537-023-00772-x.

[17] C. Su and W. Wang, "Concrete cracks detection using convolutional neural network based on transfer learning," *Journal of Electrical and Computer Engineering*, vol. 2020, Article ID 7240129, 2020. doi: 10.1155/2020/7240129.

[18] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. Al-Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021. doi: 10.3390/rs13030516.

[19] J. Maurcio, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, p. 5521, 2023. doi: 10.3390/app13095521.

[20] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jeou, "Going deeper with image transformers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 12917–12927. doi: 10.1109/iccv48922.2021.00010.

[21] K. Han *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 6743–6761, 2022. doi: 10.1109/tpami.2022.3152247.

[22] Z. Pan, B. Zhuang, H. He, J. Liu, and J. Cai, "Less is more: Pay less attention in vision transformers," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, pp. 1345–1353, 2022. doi: 10.1609/aaai.v36i2.20099.

[23] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: A fast hybrid vision transformer using structural reparameterization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 532–541. doi: 10.1109/iccv51070.2023.00532.

[24] N. Hüttén, R. Meyes, and T. Meisen, "Vision transformer in industrial visual inspection," *Applied Sciences*, vol. 12, no. 23, p. 11981, 2022. doi: 10.3390/app122311981.

[25] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, "Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images," *IEEE Access*, vol. 11, pp. 12345–12355, 2023. doi: 10.1109/ACCESS.2023.3326528.

[26] P. Wang, Q. Yang, Z. He, and Y. Yuan, "Vision transformers in multi-modal brain tumor MRI segmentation: A review," *Measurement*, vol. 210, p. 100004, 2023. doi: 10.1016/j.metrad.2023.100004.

[27] X. Li and S. Li, "Transformer help CNN see better: A lightweight hybrid apple disease identification model based on transformers," *Agriculture*, vol. 12, no. 6, p. 884, 2022. doi: 10.3390/agriculture12060884.

[28] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, vol. 11, pp. 13456–13465, 2023. doi: 10.1109/access.2023.3293813.

[29] S. Jamil, M. Abbas, and A. M. Roy, "Distinguishing malicious drones using vision transformer," *AI*, vol. 3, no. 2, p. 16, 2022. doi: 10.3390/ai3020016.

[30] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19, and ResNet50 architecture frameworks for image classification," *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/document/9687944.

[31] S. Vats, J. P. Bhati, A. Singla, V. Kukreja, and R. Sharma, "Advanced image classification on Intel datasets using optimized EfficientNet and MobileNetV2," *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10543649.

[32] A. Kaur, V. Kukreja, N. Thapliyal, M. Aeri, R. Sharma, and S. Hariharan, "Fine-tuned EfficientNet and MobileNetV2 models for Intel images classification," *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10512279.

[33]Yimin Cai1,Yuqing Long **"**Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution" doi: https://doi.org/10.1186/s12911-023-02129-z

[34]Ahmad Ishaq,Fath U Min Ullah**,"**Improved EfficientNet Architecture for Multi-Grade Brain Tumor Detection",doi:https://doi.org/10.3390/electronics14040710

[35]Ze Liu,Han Hu "Swin Transformer V2: Scaling Up Capacity and Resolution"
doi: https://arxiv.org/abs/2111.09883v2

[36]Rizka Yulvina,Mia Rizkiania **"**Hybrid Vision Transformer and Convolutional Neural Network for Multi-Class and Multi-Label Classification of Tuberculosis Anomalies-on-Chest-X-Ray"
doi:https://doi.org/10.3390/computers13120343

[37]Ze Liu,Yutong Lin**"**Swin Transformer: Hierarchical Vision Transformer using Shifted Windows",doi:https://doi.org/10.48550/arXiv.2103.14030

ISBN: 978-93-343-1044-3