Minerals Engineering 208 (2024) 108599

Contents lists available at ScienceDirect

Minerals Engineering

journal homepage: www.elsevier.com/locate/mineng

Multivariate image processing in minerals engineering with vision transformers

Xiu Liu [a], Chris Aldrich [a,b,*]

[a] Western Australian School of Mines: Minerals, Energy and Chemical Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
[b] Department of Process Engineering, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

ARTICLE INFO

Keywords:
Particle size analysis
Hydrocyclone underflow
Computer vision
Deep learning
Vision transformers
Convolutional neural networks
Explainability analysis
Attention maps

ABSTRACT

Vision transformers (ViTs) are a new class of deep learning algorithms that have recently emerged as a competitive alternative to convolutional neural networks. In this investigation, their application to two operations previously studied in the mineral processing industry is considered. These are image recognition of fines in coal particles on conveyor belts and characterisation of the particle size in the underflow of a hydrocyclone. Promising results were achieved by use of vision transformers, as they performed as well as, or better than convolutional neural networks in these image recognition problems. In addition, features extracted from the best ViT model could be used to visualise its performance and these features could also serve as a basis for nonlinear process monitoring models. Furthermore, explainability techniques such as attention maps for ViTs were implemented to better understand the ViT models, similar to techniques such as occlusion sensitivity maps used with convolutional neural networks.

1. Introduction

From early developments in the 1970s and 1980s, image-based sensor systems have become established as an important component in plant automation. Image analysis can comprise different operations of various complexity, ranging from the identification of operational conditions based on image classification (e.g. Weixing and Liangqin, 2016), through to object identification in images based on image segmentation (e.g. Ghorbani et al., 2011; Wang et al., 2016; Wang and Chen, 2016) or higher-level image understanding (e.g. Qi et al., 2020; Hu et al., 2022).

In particular, multivariate image analysis, where the images are treated as patterns associated with process conditions, have become well established in the flotation industries (Aldrich et al., 1997; Runge et al., 2007; Duchesne, 2010; Aldrich et al., 2010,2022), and are also used for example in monitoring of particulate feeds on conveyor belts in mining and metallurgy (Miranda et al., 2012; Kistner et al., 2013; Li et al., 2022; Siami et al., 2022), hydrocyclone underflows (Olivier and Aldrich, 2021; Olivier et al., 2022), ore texture classification (Yacher et al., 1986; Tessier et al., 2007; Marchetti et al., 2022; Tang et al., 2022) and furnace operations (Lu and Wen, 2021; Nagadasari and Bojja, 2022; Popov and Todeschini, 2022).

Owing to their deep architectures and large parameters sets, convolutional neural networks (CNNs) have essentially displaced traditional approaches in image recognition as the state-of-the-art in a rapidly increasing number of application scenarios. On the downside, these networks are computationally expensive and typically require high performance computing environments to train and maintain. Moreover, in some applications, CNNs' inability to adequately capture the structural dependency between its features may inhibit their scalability to more generalised image interpretation (Guo et al., 2022).

In contrast, transformers (Khan et al., 2022) are deep neural network architectures that have originated from a need to address the challenges associated with natural language processing (Worsham and Kalita, 2020; Peer et al., 2022). Like their CNN counterparts in image processing, transformers have very rapidly become the dominant architecture used in natural language processing (Han et al., 2022). Vision transformers (ViTs) are transformer architectures adapted to image processing, where instead of processing sequences of words, they process sequences of images or image patches.

Vision transformers are rapidly rising as a competitive alternative to CNNs, owing to their superior scalability to larger databases, smaller image-specific inductive biases, higher computation efficiency, their

* Corresponding author at: Western Australian School of Mines: Minerals, Energy and Chemical Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia.

E-mail address: chris.aldrich@curtin.edu.au (C. Aldrich).

https://doi.org/10.1016/j.mineng.2024.108599
Received 7 January 2023; Received in revised form 19 January 2024; Accepted 20 January 2024
Available online 3 February 2024
0892-6875/

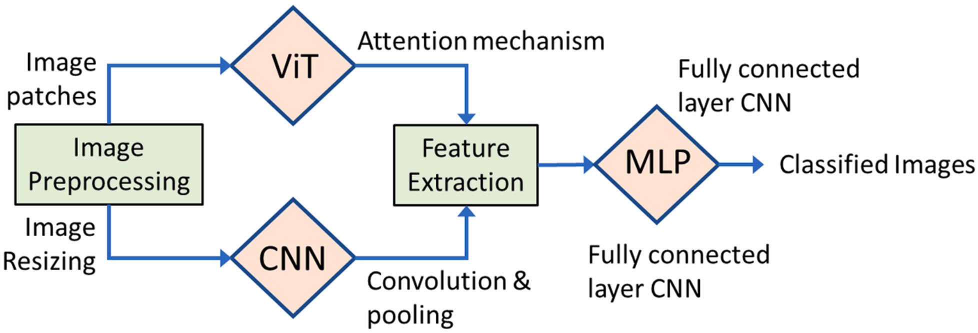
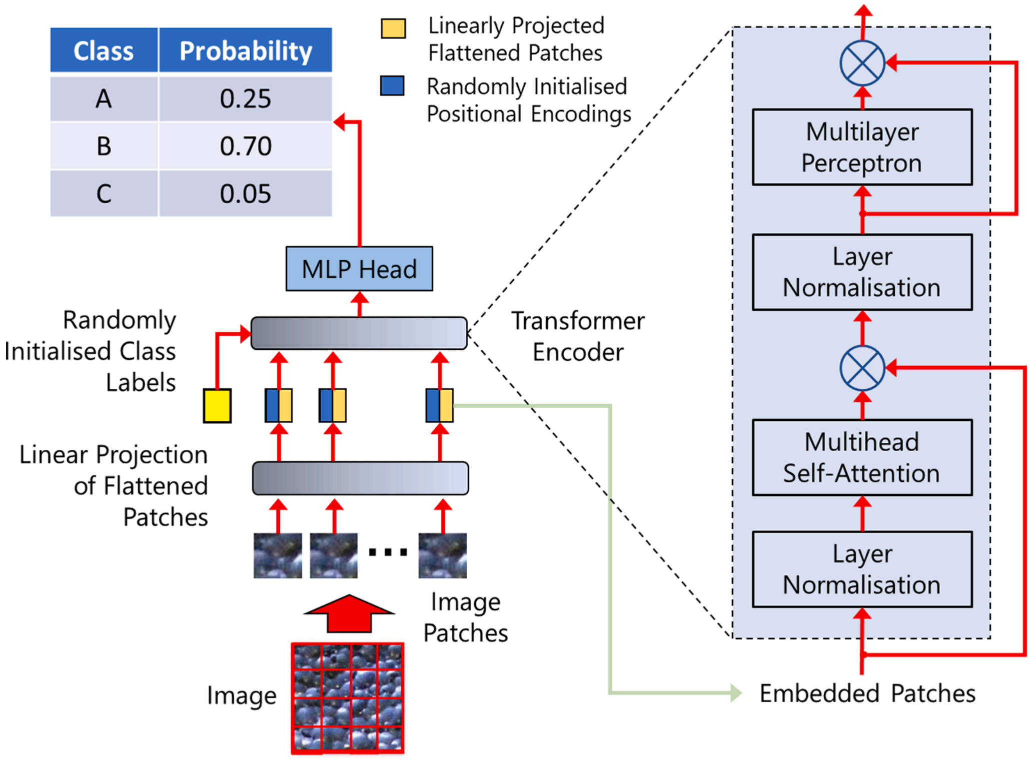


Fig. 1. Basic analytical workflow for multivariate image analysis with ViT and CNN deep learning models.



Class	Probability
A	0.25
B	0.70
C	0.05

Linearly Projected Flattened Patches
Randomly Initialised Positional Encodings
MLP Head
Randomly Initialised Class Labels
Transformer Encoder
Linear Projection of Flattened Patches
Image Patches
Image
Embedded Patches
Multilayer Perceptron
Layer Normalisation
Multihead Self-Attention
Layer Normalisation

Fig. 2. Generic image processing with a vision transformer.

ability to handle sequences of images, as well as being capable of interpreting structural dependencies between image features.

Recent work has shown that ViTs can achieve comparable or superior performance on image classification tasks at a significantly lower cost, although the results of comparative studies may vary. For example, Tuli et al. (2021) have among other shown that a ViT-B32 model was significantly more accurate than a Resnet-50 CNN model when benchmarked on variants of the ImageNet image data set. Dosovitskiy et al. (2021) have likewise found that vision transformers outperformed CNNs on various benchmark data sets, albeit by a very narrow margin. Deininger et al. (2022) have compared the DeiT-Tiny vision transformer with ResNet-18, a state-of-the-art convolutional neural network in the recognition of tumors and have found that the ViT performed slightly better than the CNN for three of four tissue types. Hütten et al. (2022) have shown that vision transformers significantly outperform CNNs on complex tasks. In contrast, Fanizzi et al. (2023) have concluded that transformers did not perform better than CNNs in predicting the recurrence of non-small cell lung cancer.

As a consequence, vision transformers are becoming well-established in different fields, where image analysis plays an important role, e.g. the health sciences (He et al., 2022) and face recognition (Luo et al., 2022).

ViTs have only very recently started to attract interest in the process industries, such as geometallurgy and mineral prospecting and mineral identification (Cui et al., 2022; Gao et al., 2024), mineral processing (Liu and Aldrich, 2023) and remote sensing (Wang et al., 2022). In some of these studies, the proposed ViT models showed better and more robust recognition than CNNs. However, very few studies have been documented to date, and the potential benefits of ViTs to image analysis in mineral processing is not well-established at present.

Therefore, in this study, the use of ViT neural networks designed to identify different concentration levels of fines in coal particles on conveyor belts, as well as the particle size analysis in hydrocyclone underflow image data is considered and compared with previous results obtained with traditional methods, as well as with deep convolutional neural networks. The results suggest that their performance can match or exceed that of traditional methods and CNNs and that ViTs are therefore a new class of deep learning methods that show exceptional promise in the development of sensors in mineral processing.

Section 2 of the paper gives a high-level overview of the analytical methodology followed in the paper, after which Sections 3 and 4 present

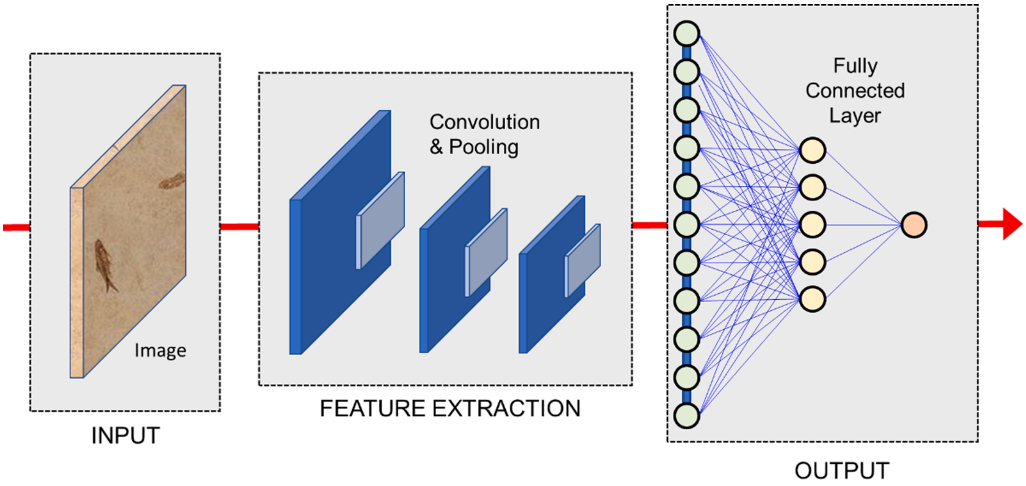


Fig. 3. Simplified architecture of a convolutional neural network.

two case studies. In Section 5, the conclusions of the paper are summarised.

2. Analytical methodology

The basic analytical steps involve the acquisition of images, preprocessing of the images, extraction of features from the images and use of these features in models capturing the relationships between the images and their labels, as indicated in Fig. 1. Preprocessing of the images could include removal of outliers (corrupted images) and resizing of the images where necessary. Data augmentation is also common, where the image database is enlarged by flipping of images, rotation or other means. Enhancement of the image contrast and centring the pixel values around zero can also be useful in scenarios where lighting conditions vary.

Both convolutional neural networks and vision transformers make use of internal feature extraction guided by their ability to predict the labels. This is a more powerful approach than traditional methods that depend on engineered features and the main reason for the success of these deep learning methods over their more traditional counterparts.

2.1. Vision transformers

With ViTs, image classification problems are cast as sequence prediction tasks for series of image patches. This enables them to capture long-term dependencies within the input image. CNNs do not have this capability. In contrast, they learn to extract features hierarchically from images, among other owing to their use of progressively enlarged receptive fields. Interested readers could refer to the seminal or review papers by among other Vaswani et al. (2017), Dosovitskiy et al. (2021), Khan et al. (2022) and Han et al. (2023).

Image processing with vision transformers broadly consists of the following steps, as outlined in Fig. 2 and briefly summarised below:

i. Splitting of images into a series of image patches of a given size

More formally, an image of size $L \times B$ with C channels, $X \in \mathbb{R}^{L \times B \times C}$ is used to generate $1, 2, \cdots N = LB/P^2$ flattened square image patches of size P that can be denoted as $X_p \in \mathbb{R}^{N \times P^2 C}$. For simplicity, image patches are square.

ii. Generating linear lower-dimensional embeddings from the flattened patches

Typically by passing through a feedforward layer with a linear activation, to generate $N \times D$ patch vectors with embedding hyperparameter D, i.e. $X_{p'} \in \mathbb{R}^{N \times D}$ To this a learnable class embedding vector is added to give vectors, $X_{p''} \in \mathbb{R}^{(N+1) \times D}$.

iii. Adding positional information to the embeddings

Positional encoding vectors are usually added to the D-dimensional patch vectors $X_{pos} \in \mathbb{R}^{(N+1) \times D}$. This gives linearly embedded patch vectors $Z = X_{p''} + X_{pos} \in \mathbb{R}^{(N+1) \times D}$.

iv. Presentation of this sequence to a standard transformer encoder
a. The encoder layer of the transformer consists of multiple encoder blocks. Each block contains a multihead attention unit and a multilayer perceptron, followed by a normalisation layer. In this block, the input vector Z is thrice duplicated and multiplied by $D \times D$ weight matrices, W_Q, W_K and W_V to obtain a query, key and value matrix, $Q \in \mathbb{R}^{(N+1) \times D}$, $K \in \mathbb{R}^{(N+1) \times D}$ and $V \in \mathbb{R}^{(N+1) \times D}$.
b. The dot product of Q and K is determined, normalized by dividing by the square root of D to obviate the vanishing gradient problem during training of the network. The normalised dot product is passed through a softmax layer and multiplied by V to produce the output or head H of the block.
c. The scaled dot products of each of the attention heads are concatenated and passed through a multilayer perceptron or dense layer to generate a final vector of embedded dimension D.
d. In the end, the Z input vector is passed through multiple such encoder blocks to produce a final context vector C. In this context vector, the context token c_0, which in the final instance is passed through a multilayer perceptron to generate the class probabilities.

As a very first introduction and application of vision transformer models in the field of mineral processing, two relatively basic or early variants of ViT models are adopted to demonstrate their capability, viz ViT-B32 and ViT-B16. These models are popular choices in many applications, where B (for 'base') is an informal descriptor of the architecture, as opposed to large (L) or huge (H), for example, while the '16' and '32' indicate the 16 × 16 and 32 × 32 patch sizes used by the models (Dosovitskyi et al., 2021). This means that ViT-B32 has a larger receptive field than ViT-B16, which can be beneficial for better capturing global features in the image. However, this comes at the cost of increased computational complexity and memory usage.

Both models are pretrained on the ImageNet-21k dataset and fine-tuned on the ImageNet-1k dataset. The ImageNet database (https://image-net.org/) consists of more than 14 million annotated

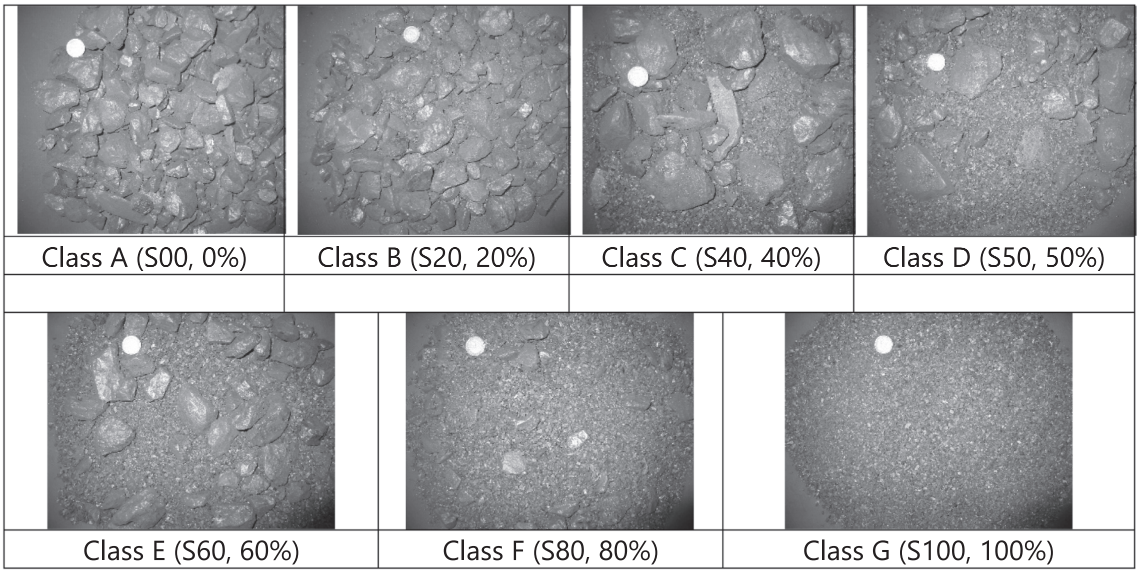


Fig. 4. Exemplars of original 2272 × 1704 images of pilot plant coal ore on a conveyor belt associated with Classes A-G considered in Case Study 1. As an indication of scale, the diameter of a bimetallic South African R5 coin shown in each image is 26 mm. The fines content of each image is indicated in parentheses.

images of 1000 common objects (Russakovsky et al., 2015), of which the above are subsets commonly used for benchmarking purposes. This image database also formed the basis of the ImageNet Large Scale Vision Recognition Challenge (https://image-net.org/challenges/LSVRC/) that was held from 2010 to 2017.

By use of the transfer learning mode, the input images were resized to 224 × 224 resolution for these two models. ViT-B32 model has 12 transformer layers, 88.3 million trained parameters and its patch size is 32 × 32 (Dosovitskyi et al., 2021). ViT-B16 model also has 12 transformer layers with slightly fewer trained parameters (86.86 million). The patch size is 16 × 16. The smaller the patch size the model has, the more resource-intensive the model is, owing to the inversely proportional relationship between the model's length sequence and the patch size square.

Moreover, it should be noted that since images presented to the pretrained vision transformers are resized to 224 × 224 pixels, this meant that with the differences in patch sizes of the two models, the ViT-B32 model produced a sequence of vectors from $(224/32)^2 = 49$ patches, and the ViT-B16 model produced a sequence of vectors from $(224/16)^2 = 196$ patches.

As can be seen from the architecture of ViT models, they effectively use the position information of the patches to greatly reduce the computational complexity. ViT has less image-specific inductive bias than CNNs and can handle arbitrary sequence lengths (subject to memory constraints).

2.2. Convolutional neural networks

The basic architecture of a convolutional neural network is shown in Fig. 3. CNN architectures are mainly composed of three consecutive sections, that is, an input section, a feature extraction section and an output section. Images at the input at a required size are passed through the whole architecture. Features which represent the images are extracted internally by use of convolution and pooling. These features are presented to the fully connected layers that act as the classifier and finally the predicted labels are obtained at the output layer.

Two popular CNN models were used in this work, namely GoogLeNet (Szegedy et al., 2015) and MobileNetV2 (Sandler et al., 2019). These networks had been pretrained in the ImageNet database and to enable transfer learning, the input images were resized to 224 × 224 resolution for each of these two models.

GoogleNet (Szegedy et al., 2015), or Inception V1, was developed by Google and its research partners. It was the winner of the ILSVRC competition in 2014. Its architecture is featured by 1 × 1 or pointwise convolutions, global average pooling, as well as inception modules. It is 22 layers deep and contains approximately 7 million trained parameters or weights.

MobileNet (Howard et al., 2017) is a relatively novel and lightweight convolutional neural network architecture adapted for use on mobile devices by significantly decreasing the number of operations and memory required, without sacrificing accuracy. Its architecture is characterised by inverted residual blocks with linear bottlenecks. It takes a low-dimensional compressed representation as input, which is subsequently expanded to a higher dimension and filtered using lightweight depth-wise convolution. Afterwards, linear convolution is used to project the features back to a low-dimensional representation. It is 53 layers deep, but contains only 3.4 million trained parameters.

3. Case Study 1: Detection of fines in particulate coal feeds

In unit operations, such as metallurgical furnaces or fluidised beds in power plants dependent on coal or coke feed material, the fines content in the feed needs to be monitored closely to prevent an excess of fines that could have a critical, adverse impact on the performance of process system (Aldrich et al., 2010).

In this case study, a dataset of coal ore piled on a pilot-plant conveyor belt, as discussed by Jemwa and Aldrich (2012), was considered here to classify the coal ore in terms of the bulk fraction of fines or fines content. The proportion of fines or fines content is defined as the percentage of particulates passing a 6 mm sieve size (% −6 mm). A total of seven classes of coal particles with varying fines compositions (0 %, 20 %, 40 %, 50 %, 60 %, 80 %, 100 %) were manually prepared by sieving a batch of industrial coal into two parts—fines and coarse—and then taking properly weighted aggregates from each part in the required ratio.

Partial simulation of the industrial conditions was realised by mixing and distributing each blended aggregate onto a pilot plant moving conveyor belt equipped with a hopper. 10 images of each mixture were captured. In this case study, each original high-resolution 2272 × 1704-pixel image was split into 16 smaller 568 × 426-pixel images. An example of each of these original images is shown in Fig. 4. The South African R5 coin with a diameter of 26 mm in each image serve as an indication of the actual particle sizes.

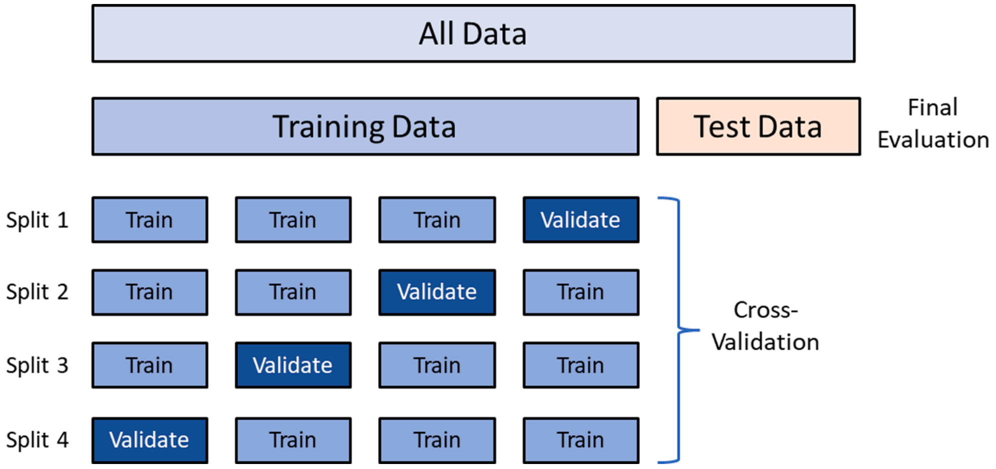


Fig. 5. Training, validation and test data used during development of the ViT models.

Table 1
Classification performance of deep learning architectures on the test image data set in Case Study 1.

Model	Number of Features	Accuracy (%)
GoogLeNet	1024	70.5
MobileNetV2	1280	71.9
ViT-B32	768	68.8
ViT-B16	768	81.3

Further analysis was restricted to these smaller images, so that as a result, each class contained 160 images and the whole dataset contained 1120 images in total. The appearance of the coal ore on the conveyor belt was cast as a textural pattern recognition problem and the results are reported based on the smaller 568 × 426-pixel images.

The two convolutional neural networks, namely GoogLeNet and MobileNetV2 and the two vision transformers, ViT-B32 and ViT-B16 were trained to classify images as belonging to one of the seven classes shown in Fig. 4.

3.1. Classification of coal particles with convolutional neural networks

GoogLeNet and MobileNetV2 were used as representative of high performance CNNs. As suggested in a previous study by the authors (Liu and Aldrich, 2022), transfer learning strategy makes a marked difference, even when dealing with small data sets, and full retraining is recommended when sufficient computational resources are available. Therefore, full retraining or fine-tuning was used, that is, all the trainable parameters in the CNNs were updated during the training process.

3.2. Classification of coal particles with vision transformers

In the second approach, coal particle images were identified from features extracted by use of two basic ViTs, viz. ViT-B32 and ViT-B16. Similarly, all the trainable parameters in the ViTs were also updated during the training process.

The convolutional neural networks used in Case Study 1 were built using a PyTorch backend, while the vision transformers were constructed using a TensorFlow backend. All the experiments were run on a GPU device on the Google Colab platform.

During retraining of the CNNs or ViTs, images of the coal particles were randomly split into training and test data sets in a ratio of 8:2, with the latter used as an independent test set to validate the generalization of the deep learning models. The training set was further randomly shuffled, and 75 % of it was used to train the models, while the remaining 25 % was allocated to a validation image data set. This was done on a four-fold basis. After training was completed, the models were tested with the test data set not used during training and validation of the models. This is illustrated in Fig. 5.

The adaptive momentum estimation (ADAM) algorithm (Kingma & Ba, 2017) was used as the optimizer in this work. Hyperparameter optimization was done by use of a grid search. Different optimal learning rates with or without an L2 penalty were applied to different CNNs or ViTs. For most models, the optimal initial learning rate was 0.0003 with a weight decay parameter of 0.00003 (L2 penalty).

The optimal batch sizes and numbers of epochs varied as well. For most models, the optimal batch size was 64 and the optimal number of epochs was 50. In order to deal with overfitting, image augmentation was used in the training stage by randomly rotating, shearing, shifting and horizontally flipping the original images. The fully retrained CNN and ViT models were used as end-to-end classifiers to discriminate between the seven classes of coal particles.

The classification performance of the different models is summarized in Table 1, together with the number of features associated with each model. The ViT-B32 model achieved reasonably good performance with an accuracy of 68.8 %. This is comparable to the results from the two fully retrained CNN models, ranging from 70.5 %~71.9 %. Furthermore, the classification accuracy with the ViT-B16 model was improved by a large margin (~10 %) to 81.3 %. Considering the much faster training speed of ViT models, ViT-B16 is the best model in Case Study 1.

The discriminative power of the different models can be further assessed by visualising the features extracted from the images by use of a *t*-distributed stochastic neighbour embedding (*t*-SNE) score plot (Van der Maaten and Hinton, 2008). The features were extracted from the "ExtractToken" layer for each of the two ViT models, as well as the layer immediately preceding the last fully connected layer for each of the two CNN models. These were the features that essentially served as the predictors of the fully connected front-ends of the networks responsible for final classification of the images. Fig. 6 shows the *t*-SNE score plots of the features extracted from the different models with the corresponding classification accuracy. In these graphs, the seven classes, S00, S20, S40, S50, S60, S80 and S100 are respectively represented by 'black dot', 'red plus', 'light blue asterisk', 'dark blue star','green triangle', 'magenta diamond' and 'yellow circle' markers.

The two fully retrained CNN features and the ViT-B32 features form seven relatively sharply delineated clusters in the feature space and thus can separate most of the seven classes very well, although class S50 seems to contribute to most of the overlap. In contrast, ViT-B16 features form more distinguishable clusters, especially for class S50, which can be further confirmed from its confusion matrix on the test set, as shown in Table 2.

It should also be noted that these results are also significantly better

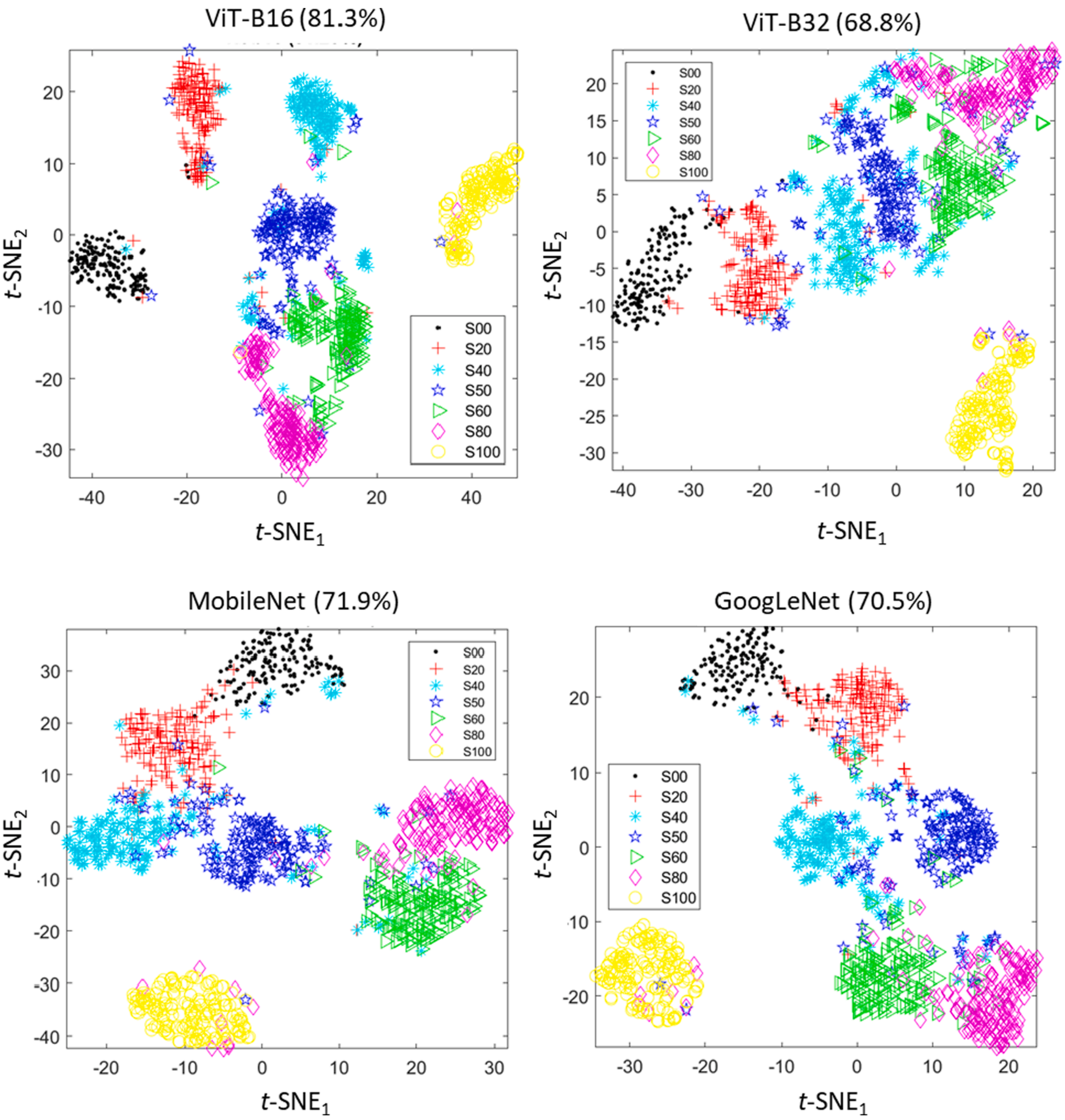


Fig. 6. *t*-SNE score plots of the features extracted from the images in Case Study 1. Top panel (left: ViT-B16, right: ViT-B32). Bottom panel (left: MobileNetV2, right: GoogLeNet). Classes S00, S20, S40, S50, S60, S80 and S100 are respectively represented by 'black dot', 'red plus', 'light blue asterisk', 'dark blue star','green triangle', 'magenta diamond' and 'yellow circle' markers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Confusion matrix on test set for fine-tuned ViT-B16 in Case Study 1.

Confusion Matrix		Predicted						
		S00	S20	S40	S50	S60	S80	S100
Actual	S00	29	2	0	0	1	0	0
	S20	2	25	1	3	1	0	0
	S40	0	1	23	4	3	1	0
	S50	2	2	2	21	4	1	0
	S60	0	0	1	3	27	1	0
	S80	0	0	1	0	2	26	3
	S100	0	0	0	0	0	1	31

than what had been achieved by Jemwa and Aldrich (2012), using optimised textons to extract features from the images that then served as predictors to support vector machines. Their best model had an overall average accuracy of 66.9 % compared to ViT-B16's 81.3 %. Moreover, the vision transformer was able to identify the 20 % class with a 78.1 % accuracy, as opposed to approximately 52 % previously documented by Jemwa and Aldrich (2012). This is significant, as the class is particularly relevant to industrial operations.

3.3. Attention maps

Finally, it is also possible to generate so-called attention maps with ViTs (Abnar and Zuidema, 2020). As the name suggests, these maps show areas in the images that are weighted more heavily in others when the image is classified. Examples of these maps using ViT-B16 model are shown in Fig. 7. The first and second columns are the class name and the original image, respectively. The third column is the attention map obtained from ViT_B16 without fine-tuning by simply passing the original image through the untrained model, and this column is presented here for better comparison. The approximate measure of scale shown on original images are in mm and this varied, as the original images were

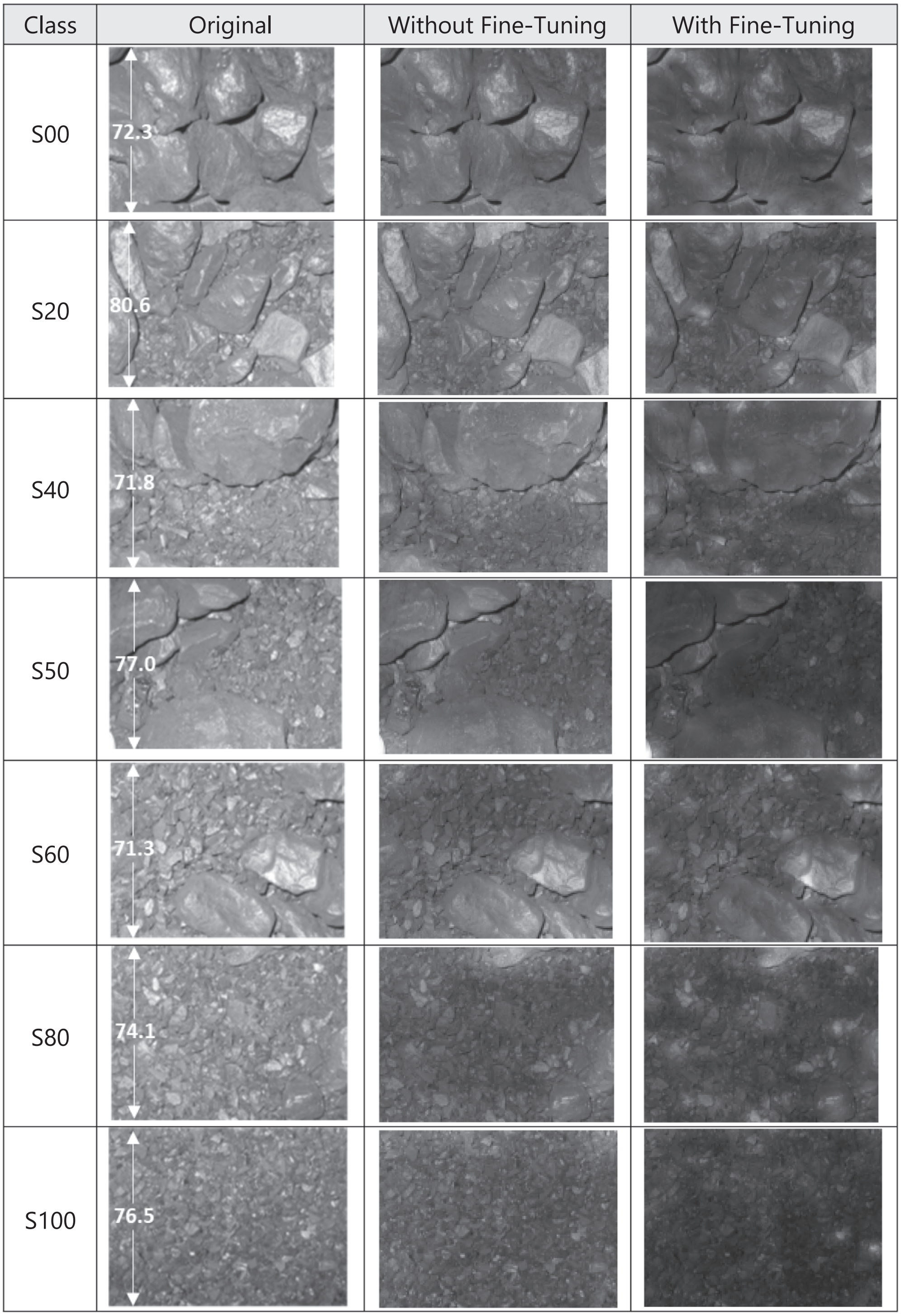


Fig. 7. Examples of attention maps of coal ore images generated with ViT-B16, showing images from classes S00 to S100 from top to bottom respectively. The left column (with scale in mm) shows the image, while the middle and right columns show the areas that the transformer focused on during classification of the images.

not taken with a completely fixed camera setup.

The last column is the attention map obtained from ViT_B16 after fine-tuning. The more heavily weighted, the brighter the specific area is. It should be noted that attention maps are not necessarily particularly informative when dealing with textural images, such as these images of coal particles, as image features may be interpreted collectively to identify patterns.

Nonetheless, as indicated by these maps, it seems as if the vision transformer took its cues from the reflections and the associated highlighted pixels, as well as the edge profiles of the coarser particles in the classification of the images. In contrast, the finer particles seemed to have received less attention. This is somewhat similar to how human eyes may see these images, that is, attention is more likely to be captured by the coarser particles with specific reflections and more pronounced edge profiles.

4. Case Study 2: Analysis of hydrocyclone underflow streams

In the second case study, a dataset of hydrocyclone underflow images collected from experiments on a laboratory hydrocyclone setup at

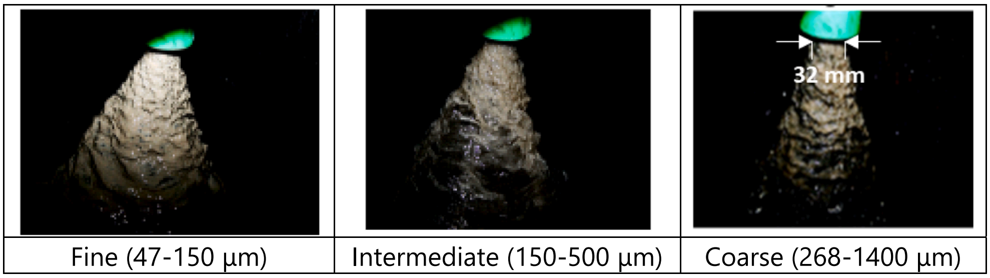


Fig. 8. Typical images of the hydrocyclone underflow of the PGM slurries associated with the three classes (from left to right: Fine, Intermediate and Coarse) in Case Study 2. The size range for each class is indicated in parenthesis. The 32 mm spigot diameter of the hydrocyclone, as indicated in the panel on the right in the figure, serve as a measure of scale.

Table 3
Comparison of deep learning models on the test set of images in Case Study 2.

Model	Number of Features	Accuracy (%)
GoogLeNet	1024	92.2
ViT-B32	768	93.3
ViT-B16	768	96.7

Stellenbosch University in South Africa was considered. These data obtained from platinum metal group ores obtained from the Merensky, Platreef and UG2 reefs in the Bushveld Igneous Complex in South Africa are discussed by Aldrich et al. (2015) and Olivier and Aldrich (2021) and are revisited in this case study, focusing on classification of the mean particle sizes in the underflow.

During the experiments, both streams of underflow and overflow are fed back into a mixing tank, creating a closed-circuit slurry flow. Varying particle size distributions in the underflow were obtained by conducting experiments at different solids loadings in the mixing tank. The system reached stabilized after each increase in the solids loading, and then the underflow images were collected.

Simultaneously, samples of the overflow and underflow streams were collected, and each sample's particle size distribution determined using sieve analysis and a Saturn DigiSizer laser particle size analyser. 15 such experiments yielded a total of 300 images of the hydrocyclone underflow with mean particle sizes ranging from 47 to 1400 µm. The images were grouped into three classes (Fine, Intermediate and Coarse) based on the mean particle size measured. Examples of these images are shown in Fig. 8. The number of images associated with the Fine, Intermediate and Coarse classes were 100, 40 and 160, respectively.

The same framework for classification of the different classes as in Case Study 1, was used in Case Study 2. The same training procedure was followed as in the first case study. The only difference was that the split ratio of training and test sets was 0.7:0.3 in Case Study 2. The

Table 4
Confusion matrix on test set for fine-tuned ViT-B16 in Case Study 2.

Confusion Matrix		Predicted		
		Intermediate	Fine	Coarse
Actual	Intermediate	11	0	1
	Fine	2	28	0
	Coarse	0	0	48

Table 5
Confusion matrix on test set for fine-tuned GoogLeNet in Case Study 2.

Confusion Matrix		Predicted		
		Intermediate	Fine	Coarse
Actual	Intermediate	7	1	4
	Fine	0	30	0
	Coarse	1	1	46

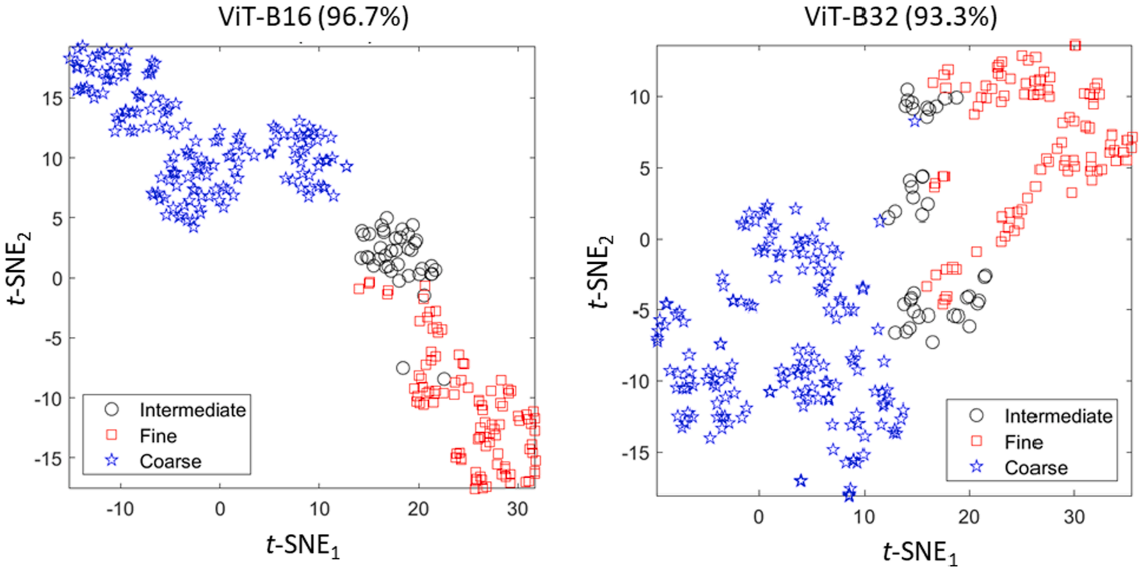


Fig. 9. Visualisation of the features generated by the ViT-B16 (left) and ViT-B32 (right) from images of the hydrocyclone underflow considered in Case Study 2. Fine, intermediate and coarse particles are indicated by red squares, black circles and blue star markers respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

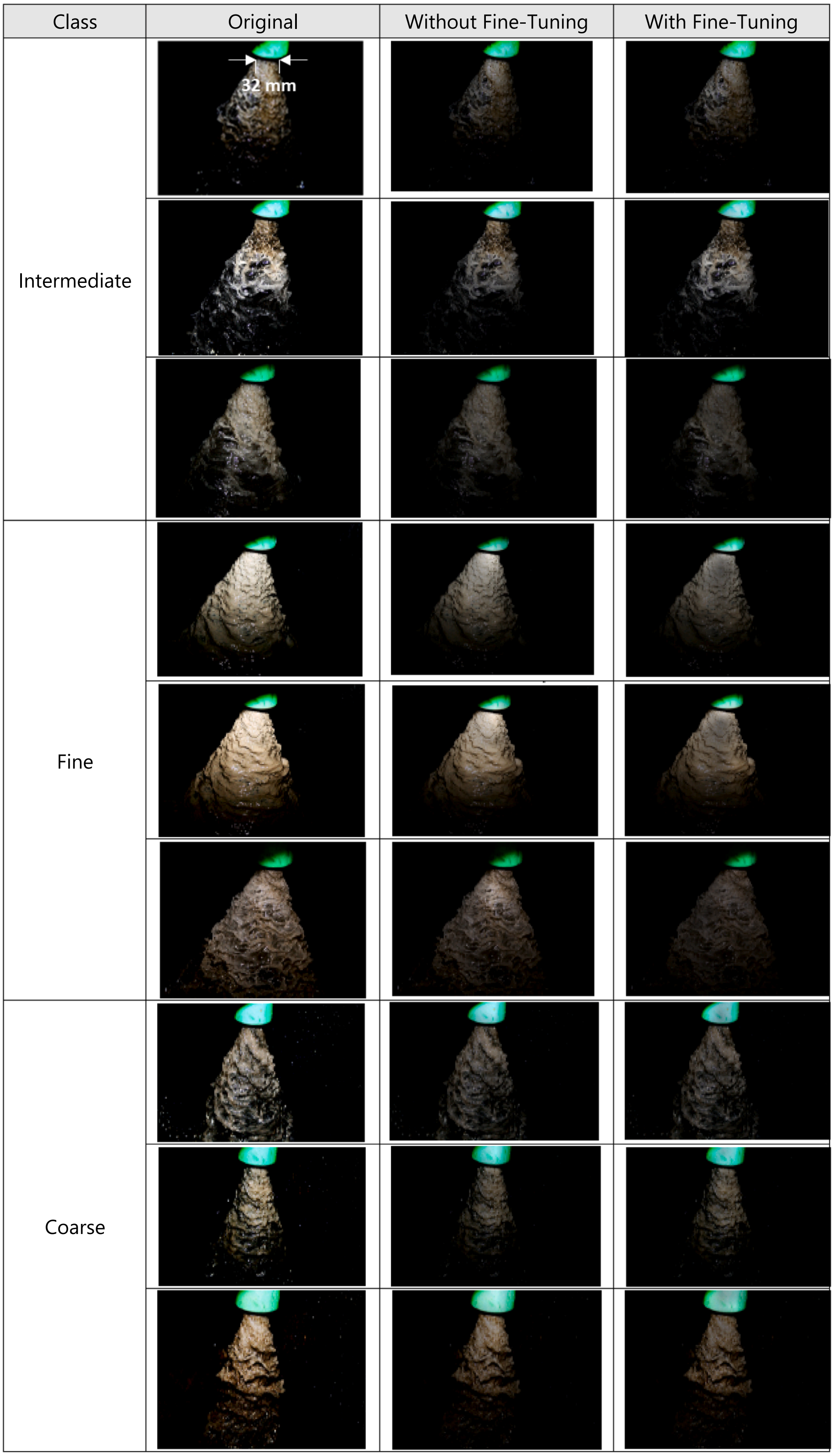


Fig. 10. Examples of attention maps of hydrocyclone underflow images generated with ViT-B16 for the PGM slurries considered in Case Study 2. The left column shows the original image. Shaded parts in the images in the middle and right columns indicate the parts of the images that played the most important role in their classification. As indicated in the top left panel, the 32 mm spigot diameter of the hydrocyclone can be used as an approximate measure of the scale of the images.

classification performance of the different models used in Case Study 2, is summarised in Table 3, together with the dimension of the corresponding feature set. The results associated with GoogLeNet are taken from the work by Olivier et al. (2022). Again ViT-B32 achieved comparable performance (93.3 %) to GoogLeNet (92.2 %) and the best performance was achieved by ViT-B16 with an accuracy of 96.7 %. As before, the features extracted by different models were visualised by *t*-SNE score plots as shown in Fig. 9. As the features extracted from GoogLeNet are not directly available, the confusion matrices (Table 4 and Table 5) were used for comparison instead.

As can be seen from the *t*-SNE plots (Fig. 9) and confusion matrices, ViT-B16 in is again the best model, as it can perfectly discriminate Class 'Coarse' from the other two classes and show reduced misclassification of Class 'Intermediate' to Class 'Coarse'.

Attention maps were also obtained for three examples of each class, as shown in Fig. 10. Although not as clearly as in Case Study 1, one can still see the differences by a careful look at which area in the image is brighter or dimmer. It seems that some areas within the texture of the underflow is brighter for Class 'Intermediate', while dimmer for Class 'Fine'. In contrast, all the areas of the textural part of the underflow for Class 'Coarse' seem to receive the same attention, while the spattering particles receive less attention.

These observations from the attention maps of ViT models are to a large extent consistent with observations based on occlusion sensitivity maps of CNN models (Olivier et al., 2022). In the study by Olivier et al. (2022), the CNN model took its cues from the texture of the spray flow patterns for fine and intermediate particle sizes, while possibly also taking the spattering of particles and the spray angle of the underflow into account for coarse particles.

5. Conclusions

The following conclusions can be made regarding two case studies where convolutional neural networks and vision transformers were used to classify images of particles on conveyor belts and the underflow of hydrocyclones:

- Vision transformers are comparable or superior to state-of-the-art convolutional neural networks in terms of accuracies of image recognition, as well as the time it took to construct the models. These results suggest that vision transformers dealing with sequences of image patches can be considered at least as a viable alternative or possibly as a better option to convolutional neural networks in image recognition.
- The architecture of the ViT models make a difference, i.e. the smaller the image patch size presented to the ViT model, the better, but there is a computational cost trade-off.
- ViT-B16 performed the best among the models considered in the two case studies. More advanced ViT models or variants can be explored in future work.
- The features extracted from ViT models, or ViT features, can be used as such in other potential applications, such as unsupervised monitoring. In these case studies, they could be used to visualise the performance of the models, after dimensionality reduction.
- Explainability analysis is an important approach interpret the results generated with deep learning models. Explaining the ViT models with attention maps is reliable and consistent with those explaining techniques applied together with CNN models.

CRediT authorship contribution statement

Xiu Liu: Conceptualization; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Roles/Writing - review & editing. **Chris Aldrich**: Conceptualization; Data curation; Investigation; Methodology; Project administration; Resources; Supervision; Roles/ Writing - original draft; and Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors acknowledge the funding support from the Australian Research Council for the ARC Centre of Excellence for Enabling Eco-Efficient Beneficiation of Minerals, Grant Number CE200100009.

References

Abnar, S., Zuidema, W., 2020. Quantifying attention flow in transformers. arXiv: 2005.00928. https://doi.org/10.48550/arXiv.2005.00928.

Aldrich, C., Moolman, D.W., Harris, M.C., Bunkell, S.-J., Theron, D.A., 1997. Relationship between surface froth features and process conditions in the batch flotation of a sulphide ore. Miner. Eng. 10 (11), 272–281. https://doi.org/10.1016/S0892-6875(97)00107-6.

Aldrich, C., Marais, C., Shean, B.J., Cilliers, J.J., 2010a. Online monitoring and control of froth flotation systems with machine vision: a review. Int. J. Miner. Process. 96 (1–4), 1–13. https://doi.org/10.1016/j.minpro.2010.04.005.

Aldrich, C., Jemwa, G.T., Van Dyk, J.C., Keyser, M.J., 2010b. Online analysis of coal on a conveyor belt by use of machine vision and kernel methods. Int. J. Coal Prep. Util. 30, 331–348. https://doi.org/10.1080/19392699.2010.517486.

Aldrich, C., Uahengo, F.D.L., Kistner, M., 2015. Particle size estimation in hydrocyclone underflow streams by use of multivariate image analysis. Miner. Eng. 70, 14–19. https://doi.org/10.1016/j.mineng.2014.08.018.

Aldrich, C., Avelar, E., Liu, X., 2022. Recent advances in flotation froth image analysis. Miner. Eng. 188, 107823 https://doi.org/10.1016/j.mineng.2022.107823.

Cui, X., Peng, C., Yang, H., 2022. Intelligent mineral identification and classification based on vision transformer. In: Proceedings of the 9th International Conference on Dependable Systems and Their Applications, DSA 2022, 670–676. https://doi.org/10.1109/DSA56465.2022.00095.

Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., Korski, K. and Gaire, F. 2022. A comparative study between vision transformers and CNNs in digital pathology. arXiv:2206.00389 [eess.IV], https://doi.org/10.48550/arXiv.2206.00389.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16 x 16 words: Transformers for image recognition at scale. arXiv: 2010.11929. https://arxiv.org/abs/2010.11929.

Duchesne, C., 2010. Multivariate image analysis in mineral processing. In: Sbarbaro, D., del Villar, R. (Eds.), Advanced Control and Supervision of Mineral Processing Plants. Springer, London.

Fanizzi, A., Fadda, F., Comes, M.C., 2023. Comparison between vision transformers and convolutional neural networks to predict non-small lung cancer recurrence. Sci. Rep. 13, 20605 (2023). https://doi.org/10.1038/s41598-023-48004-9.

Gao, Q., Long, T., Zhou, Z., 2024. Mineral identification based on natural feature-oriented image processing and multi-label image classification. Expert Syst. Appl. 238, 122111 https://doi.org/10.1016/j.eswa.2023.122111.

Ghorbani, Y., Becker, M., Petersen, J., Morar, S.H., Mainza, A., Franzidis, J.-P., 2011. Use of X-ray computed tomography to investigate crack distribution and mineral dissemination in sphalerite ore particles. ISSN 0892-6875 Min. Eng. 24(12), 1249–1257. https://doi.org/10.1016/j.mineng.2011.04.008.

Guo, J., Jia, N., Bai, J., 2022. Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image. Sci. Rep. 12, 15473. https://doi.org/10.1038/s41598-022-19831-z.

Han, K., Wang, Y., Chen, H., et al., 2023. A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861.

Hu, J., Kong, J., Zhang, Q., Liu, R., 2022. Enhancing scene understanding based on deep learning for end-to-end autonomous driving. ISSN 0952-1976 Eng. Appl. Artif. Intell. 116.

Hütten, N., Meyes, R., Meisen, T., 2022. Vision transformer in industrial visual inspection. Appl. Sci. 12 (23), 11981. https://doi.org/10.3390/app122311981.

Jemwa, G.T., Aldrich, C., 2012. Estimating size fractions of coal particles on conveyor belts using image texture modelling methods. Expert Syst. Appl. 39, 7947–7960. https://doi.org/10.1016/j.eswa.2012.01.104.

Khan, S., Nazeer, M., Hayat, M., Zamir, S.W., 2022. Transformers in vision: a survey. Art 200 ACM Computing Surveys 54 (10s), 1–41.

Kistner, M., Jemwa, G.T., Aldrich, C., 2013. Monitoring of mineral processing systems by using textural image analysis. Miner. Eng. 52, 169–177. https://doi.org/10.1016/j.mineng.2013.05.022.

Li, M., Wang, X., Yao, H., Saxén, H., and Yu, Y. 2022. Analysis of particle size distribution of coke on blast furnace belt using object detection. Processes, 10(10), art. no. 1902. https://doi.org/10.3390/pr10101902.

Liu, X., Aldrich, C., 2022. Deep learning approaches to image texture analysis in material processing. art. No. 355 Metals 12. https://doi.org/10.3390/met12020355.

Liu, X., Aldrich, C., 2023. Flotation froth image recognition using vision transformers. IFAC-PapersOnLine 56 (2), 2329–2334. https://doi.org/10.1016/j.ifacol.2023.10.1202.

Lu, S.-W., Wen, Y.-X., 2021. Semi-supervised classification of semi-molten working condition of fused magnesium furnace based on image and current features. Zidonghua Xuebao/acta Automatica Sinica 47 (4), 891–902.

Marchetti, M., Fongaro, L., Bulgheroni, A., Wallenius, M., Mayer, K., 2022. Classification of uranium ore concentrates applying support vector machine to spectrophotometric and textural features. art. no. 105443 Appl. Geochem. 146. https://doi.org/10.1016/j.apgeochem.2022.105443.

Miranda, R.C., Martins, M.A.S., Gontijo, M.M. and Nogueira, A. 2012. Crushing and autogenous grinding advanced control assisted by image analysis. In: 26th International Mineral Processing Congress (IMPC 2012), 3433–3438, New Delhi, India, 24–28 Sep.

Nagadasari, M.P., Bojja, P., 2022. Industrial IoT enabled fuzzy logic based flame image processing for rotary kiln control. Wirel. Pers. Commun. 125 (3), 2647–2665. https://doi.org/10.1007/s11277-022-09677-z.

Olivier, J., Aldrich, C., 2021. Underflow particle size estimation of hydrocyclones by use of transfer learning with convolutional neural networks. IFAC-PapersOnLine 54 (11), 85–90. https://doi.org/10.1016/j.ifacol.2021.10.055.

Olivier, J., Aldrich, C., Liu, X., 2022. Explaining convolutional neural network predictions of particle size in the underflow of a hydrocyclone. IFAC-PapersOnLine 55 (21), 19–24. https://doi.org/10.1016/j.ifacol.2022.09.237.

Popov, I., Todeschini, G., 2022. Flame intensity analysis for hot molten metal pouring in the steel industry by applying image segmentation. In: Lecture Notes in Networks and Systems. https://doi.org/10.1007/978-3-030-90532-3_47.

Qi, X., Zhu, P., Wang, Y., et al., 2020. MLRSNet: a multi-label high spatial resolution remote sensing dataset for semantic scene understanding. ISSN 0924-2716 ISPRS J. Photogram. Rem. Sens. 169, 337–350. https://doi.org/10.1016/j.isprsjprs.2020.09.020.

Runge, K., McMaster, J., Wortley, M.G., Rosa, D.L., Guyot, O., 2007. A correlation between Visiofroth(TM) measurements and the performance of a flotation cell. In: Ninth Mill Operators' Conference 2007, Australasian Institute of Mining and Metallurgy, Fremantle, WA, Australia.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) 115 (3), 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Sandler, M., Howard, A., Zhu, M., Zhmoginov A., Chen, L.-C., 2019. MobileNetV2: Inverted residuals and linear bottlenecks. arXiv:1801.04381v4 [cs.CV], https://doi.org/10.48550/arXiv.1801.04381.

Siami, M., Barszcz, T., Wodecki, J., Zimroz, R., 2022. Design of an infrared image processing pipeline for robotic inspection of conveyor systems in opencast mining sites. art. no. 6771 Energies 15(18). https://doi.org/10.3390/en15186771.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinowitz, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

Tang, K., Wang, Y.D., Mostaghimi, P., Knackstedt, M., Hargrave, C., Armstrong, R.T., 2022. Deep convolutional neural network for 3D mineral identification and liberation analysis. Min. Eng. 183, art. no. 107592. https://doi.org/10.1016/j.mineng.2022.107592.

Tessier, J., Duchesne, C., Bartolacci, G., 2007. A machine vision approach to on-line estimation of run-of-mine ore composition on conveyor belts. Miner. Eng. 20 (12), 1129–1144. https://doi.org/10.1016/j.mineng.2007.04.009.

Tuli, S. Dasgupta, I., Grant, E., Griffiths, T.L., 2021. Are convolutional neural networks or transformers more like human vision? arXiv:2105.07197 [cs.CV], https://doi.org/10.48550/arXiv.2105.07197.

Van der Maaten, L., Hinton, G.E., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Vaswani, A., Shazeer, N, Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762.

Wang, W., Chen, L., 2016. Mineral froth image classification and segmentation. In: Radhakrishnan, S. (Ed.), Recent Advances in Image and Video Coding, https://doi.org/10.5772/65028, ISBN 978-953-51-2776-5, InTechOpen.

Wang, Y., Lin, C.L., Miller, J.D., 2016. 3D image segmentation for analysis of multisize particles in a packed particle bed. ISSN 0032-5910 Powder Technol. 301, 160–168. https://doi.org/10.1016/j.powtec.2016.05.012.

Wang, A., Xing, S., Zhao, Y., Wu, H., Iwahori, Y., 2022. A hyperspectral image classification method based on adaptive spectral spatial kernel combined with improved vision transformer. art. no. 3705 Remote Sens. (Basel) 14(15). https://doi.org/10.3390/rs14153705.

Yacher, L., Mujica, L.F., Gonzalez, C. and Nobile, R. 1986. Industrial trials for an image coarse particle analyzer in a SAG mill. Preprint - Society of Mining Engineers of AIME, 6 p.