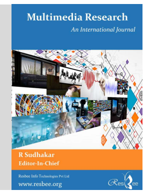
Vol.7 No.4 Oct 2024

Multimedia Research

Received 18 October, Revised 23 October, Accepted 8 November




# Terrain Classification using Deep Learning

**Gurugubelli Anil**
*BTech, Computer Science Engineering*
*GMR Institute of Technology*
*gurugubellianil21@gmail.com*

**Abstract:** Terrain refers to a geographical tract of land over the surface of the Earth. Terrain recognition refers to the ability to identify and analyze the characteristics of a given terrain. Our project involves the process of classifying the terrain into different types and recognizing a new image of a piece of land as part of its corresponding terrain class. Furthermore, we analyze the texture of the terrain and compute the levels of roughness and slipperiness of the specific terrain. The main objective of this task is to enhance high-level environment perception to gain deeper insights into the classified terrain. For this purpose, we employ Machine Learning and Deep Learning techniques. Terrain recognition and classification requires data composed of images of various terrain types, which is further pre-processed. These pre-processed images undergo feature extraction using pre-trained models such as VGG-19, ResNet-50, and Efficient Net. The extracted features are vectorized to convert them into numeric values, making the training process more efficient. The model is trained for classification using Deep Learning models like Vision Transformers, CNN, and Lidars. To compute the implicit qualities of the terrain and perform surface texture analysis, we employ algorithms such as the Gray-Level Co-Occurrence Matrix (GLCM) and Gabor filters. As a result, the terrain is recognized along with its implicit qualities, providing diverse environmental perceptions. This task of terrain recognition is important and highly beneficial in fields such as the military, navigation, and agriculture.

**Keywords**: Terrain, Environmental Perception, VGG-19, ResNet-50, Efficient Net, Vision Transformers, CNN, Lidars, Gray- Level Co-Occurrence Matrix, Gabor filters

***Nomenclature***

| Abbreviation | Descriptions |
|---|---|
| **CNN** | Convolutional Neural Network |
| **GLCM** | Gray-Level Co-Occurrence Matrix |
| **VGG-19** | Visual Geometry Group (19-layer) |
| **ResNet-50** | Residual Network (50-layer) |
| **Lidar** | Light Detection and Ranging |
| **EfficientNet** | Efficient Neural Network |

## 1. Introduction

Terrain Recognition is crucial for various applications such as robotics, autonomous vehicles, military operations, and environmental monitoring. Its primary goal is to categorize landforms into four major classes: Sandy, Rocky, Marshy, and Grassy. This task leverages advanced Deep Learning models, including Vision Transformers, Residual Networks, and Encoders, which have proven highly effective for terrain classification. The process begins with training the model on a comprehensive image dataset that showcases diverse landforms. This dataset undergoes several preprocessing techniques to enhance the quality and usefulness of the data. One key technique is segmentation, which involves dividing the images into meaningful regions or objects. This step facilitates easier data analysis and processing by isolating relevant features within the images. Another crucial preprocessing method is super-resolution, which improves the image quality by increasing the pixel count and enhancing visual clarity. This allows the model to work with higher-resolution images, capturing finer details that are essential for accurate classification. Once preprocessing is complete, the images are tokenized into smaller patches, with each patch treated as an individual token. These tokens serve as inputs to the transformer model, which is designed with multiple layers to handle positional encoding. Within the transformer model, self-attention layers play a pivotal role in capturing the relationships between different patches. This mechanism enables the model to understand the global context of the image, which is critical for accurate terrain classification. The training process involves fine-tuning the model to achieve optimal performance. Once trained, the model is capable of accurately classifying new terrains into their respective categories. This capability allows for real-time recognition of landforms, which provides valuable insights for decision-making in various environments. For instance, in robotics, this technology enables robots to navigate and operate more efficiently by understanding the terrain they are traversing. In military operations, accurate terrain recognition can significantly enhance strategic planning and execution by providing detailed information about the battlefield environment. Environmental monitoring also benefits greatly, as this technology aids in the assessment and management of natural landscapes. Overall, Terrain Recognition, powered by deep learning approaches, emerges as a highly convenient and crucial tool for understanding and navigating diverse terrains. Its ability to process and analyze complex landform data in real-time offers significant advantages across multiple fields, making it an indispensable asset for modern technological applications. The rest of the paper is organized as: Section 2 covers the literature survey, Section 2 explains the methodology, Section 3 details the result and discussion, Section 4 mentions the advantages and disadvantages, and Section 6 concludes the paper.

## 2. Literature Survey

In 2023, Yang, F., et al. [1] created a breakthrough that occurred with the development of a depth camera-based Vision Transformer (ViT) algorithm tailored for wearable soft exosuits. This innovation focused on precise terrain recognition without providing power assistance for stairs or similar environments, addressing the need for adaptability to complex terrains. The algorithm's accuracy stems from its ability to efficiently process depth data, leveraging ViT's strengths in capturing spatial dependencies and hierarchical features. By analyzing intricate terrain features, including slopes, uneven surfaces, and obstacles, the system enables wearers to navigate diverse environments with confidence and stability. Its effectiveness lies in its capacity to adapt to varying terrains seamlessly, enhancing user safety and mobility. This advancement represents a significant step forward in the integration of computer vision and wearable technology, promising enhanced functionality and usability for soft exosuits in real-world scenarios.

In 2023, Moore., et al. [2] aimed to enhance free-living fall risk assessment through data mining and deep learning techniques. The focus was on improving contemporary mobility analysis by accurately classifying indoor and outdoor environments and terrains. Despite notable achievements such as an F1 Score of 0.89 and 80% accuracy, limitations were evident. These included the inability to provide a broader context, reliance on bulky computational setups, and challenges in acquiring data across various conditions like lighting or weather. However, the study proposed automated computing methods employing Artificial Intelligence to overcome these challenges. By providing a comprehensive understanding of environmental context, particularly through IMU-based mobility analysis, the model aimed to reduce fall risk and enhance diagnostic processes. Furthermore, the incorporation of accelerometer-based measurements for head angle determination aimed to refine AI algorithms, thereby improving accuracy in fall risk assessment and decision-making processes.

In 2022, Iwahashi, et al., [4] created a global polygon for terrain classification marking a significant step forward in disaster prevention, environmental assessment, and seismic hazard mapping. By merging regions with similar topographic characteristics, this approach aimed to calculate terrain measurements more accurately. Challenges included potential spatial arrangement disruptions, difficulty in handling noisy areas, and limitations in representing changes due to human activities. Nonetheless, this method improved the classification of plains by utilizing finer Digital Elevation Models (DEMs) and efficiently classifying small and forms. Additionally, segmenting and connecting polygons enhanced accuracy and reliability while overcoming the trade-off between higher-resolution DEMs and global geomorphological legend creation. Integrating remote sensing images and geological data facilitated a focus on dynamic changes in pology, providing a comprehensive understanding of terrain characteristics. This interdisciplinary approach holds promise for better-informed disaster management strategies and environmental assessments on a global scale.

In 2023, Ramanathan, et al., [5] developed a heuristic vision-based terrain recognition for lower limb exoskeletons aimed to aid patients with reduced mobility from conditions like stroke or spinal cord injuries. This system focused on identifying terrain features through offline and online experiments. Challenges included limited accuracy and the loss of spatial information during preprocessing, as well as potential confusion caused by artifacts like the user's leg appearing in images. However, it outperformed traditional methods by recognizing more terrain classes, including obstacles and gaps, and handling scenarios where the ground is not visible. With an accuracy of 94.85%, it showed promise for integration into real-time scenarios. Future improvements include expanding the number of terrain classes for everyday use, integrating additional sensors like RGB for enhanced robustness, and conducting quantitative analyses to further refine performance. This innovation holds significant potential for improving mobility assistance technology and enhancing the quality of life for individuals with mobility impairments.

In 2023, Li, et al., [13] presented an enhanced DEM super-resolution Transformer (DSRT) network aimed at improving the resolution of large-scale Digital Elevation Models (DEMs) for geographic information. While the model's performance might be compromised in scenarios with inadequate data or suboptimal optimization, the DSRT network demonstrated superiority over conventional spatial interpolation methods. The Root Mean Square Error (RMSE) is reported as 87.466%, showcasing the effectiveness of the DSRT network in enhancing the resolution of DEMs. This research contributes to advancements in geographical information processing, offering a sophisticated solution for improving the resolution of large-scale DEMs. Further investigations into optimizing the model's performance under varying data conditions could enhance its practical applicability in real-world scenarios.

In 2023, Jiang, et al., [15] introduced an approach for super-resolution terrain modeling in High Mountain Asia (HMA), incorporating a new loss function that considers terrain parameters such as slope and curvature to enhance data accuracy. Despite the potential benefits of incorporating terrain parameters, challenges were noted regarding the sensitivity of curvature to errors and the scale-dependency of the slope, which could impact the accuracy of generated terrain data. Despite these challenges, the method holds promise for application in inaccessible regions on Earth and other planets with terrain resembling HMA, offering avenues for advancing scientific research across diverse geographical contexts. The method's performance was reported with a 32.17% improvement in terms of Root Mean Square Error (RMSE) accuracy and a 33.97% improvement in terms of Mean Absolute Error (MAE) accuracy, indicating its potential effectiveness in high-resolution terrain modeling tasks. Continued research could address challenges related to terrain parameter sensitivity and scalability, further enhancing the applicability and accuracy of super-resolution terrain modeling techniques.

In 2013, Lin, et al., [20] developed an effective approach for selecting terrain modeling methods was proposed, utilizing a support vector machine (SVM) classifier to categorize terrain surfaces as either complex or moderate based on a terrain complexity index (TCI) associated with the terrain elevation range. However, the study encountered limitations due to a relatively small sample size for training and testing the SVM classifier, comprising 20 positive and 40 negative samples in the training set, and 7 positive and 13 negative samples in the testing set. This small sample size may compromise the robustness and reliability of the classification results. Despite this limitation, the use of TCI and SVM classifiers demonstrated an accuracy of 84.3%, enhancing the accuracy and generality of terrain type classification. This approach holds the potential for improving the selection of terrain modeling algorithms, although future research may benefit from larger and more diverse datasets to ensure the reliability of classification outcomes.

In 2023, Ewing., et al., [23] explored an alternative method for terrain characterization through the utilization of remote sensing data and machine learning algorithms. It focused on predicting soil properties like moisture and terrain strength by leveraging thermal, multispectral, and hyperspectral remote sensing data. However, the study's limitation lies in its reliance on thermal emissivity measurements from soils, restricting its application to offline scenarios. An online application for predicting soil strength would enhance practical usability. Nonetheless, the paper underscores the advantages of this approach, which include rapidity, cost-efficiency, and safety compared to traditional in-situ measurement methods. By utilizing remote sensing data and machine learning algorithms, this methodology offers a promising avenue for terrain characterization, potentially revolutionizing the way such tasks are conducted in various fields, including agriculture, geology, and environmental monitoring. Further research could focus on overcoming the offline limitation to broaden its applicability.

## 2.1 Review Table

***Table 1.*** *Review of Existing Methods.*

| Author | Objectives | Limitations | Advantages | Metrics |
|---|---|---|---|---|
| **Yang, F., *et al.* [1]** | Develop a depth camera-based ViT algorithm for accurate terrain recognition in wearable soft exosuits | Does not provide power assistance for stairs or other environments | Adapting to complex terrains | Accuracy |
| **Moore., *et al.* [2]** | Enhance free livingrisk fall assessment. Contemporary mobility analysis Accurately classify both indoor and outdoor environment and terrain. | Relies on bulky computational setup. Acquiring data over wide range of exemplars or conditions such as lightning or rain etc is difficult. To head mounted cameras and instances where the wearer turns their head, Removing the target terrain from the view. | Automated computing methods with the use of Artificial Intelligence. Provides concrete and through understanding of environmental context. Free-living IMU based mobility and lesser fall-risk. | F1 Score –0.89, Accuracy 80% |
| **Iwahashi, et al.,** | Terrain classification for | Can affect the spatial | Classify small landforms | Accuracy |

| | | | | |
|---|---|---|---|---|
| **[4]** | disaster prevention, environmental assessment, seismic hazard mapping. Merging the regions with similar topographic characteristics for calculating terrain measurements. For better Classification of plains and with stand use of finer DEM | arrangement, can't work accurately with areas under Noise and removing the noise is another big task. May not represent the change under human activities. | more efficiently. Segmenting & connecting polygons to improve accuracy and reliability. Overcomes trade-of between DEMs of higher resolution and the creation of global geomorphological legends | |
| **Ramanathan, *et al.,* [5]** | Aim to help patients who have reduced mobility due to various factors such as stroke and spinalcord injuries. Terrain recognition Including the identification of its features. Validated with offline and online experiments. | Trained to detect only a limited number of terrain classes and have limited accuracy to work with actual exoskeletons. Actual spatial information is Lost during preprocessing. based on the position of the camera, artefacts such as the user's leg and foot might appear on the image, which could confuse the terrain recognition. | Recognizes more terrain classes than traditional methods, including obstacles and gaps enabling better terrain traversal. Handle the cases when the ground is not visible. Can be integrated with real time scenarios. | Accuracy 94.85 |
| **Li, *et al.,* [13]** | The objective of the paper is to propose an Improved DEM super- resolution Transformer (DSRT) network for large- scale geographic information. | The overall performance of the model could be compromised in scenarios with insufficient dataor improper optimization. | The proposed DEM super-resolution Transformer network (DSRT) outperforms conventional spatial interpolation methods. | RMSE 87.466% |
| Jiang, et al., **[15]** | The proposed method incorporates a new loss function that considers terrain parameters such as slope and curvature to improve the accuracy of the generated terrain data. | The proposed method incorporates terrain parameters such as slope and curvature as constraints, but the sensitivity of curvature to errors and the scale dependency of slope may affect the accuracy of the generated terrain data. | The proposed method shows promise for application in other inaccessible areas on Earthor even other planets with terrain similar to HMA, offering opportunities for advancing scientific research in various geographical contexts. | 32.17% in Term of RMSE accuracy and by 33.97% in terms of MAE accuracy. |
| **Lin, [20] *et al.,*** | A terrain structure- based directed graph model (DPN) to model overall landforms as graph data with geo-meaning is introduced. We then develop a graph DL technique flow based on a graph attention network (GAT) to leverage DPN to achieve graph-driven landform recognition | The paper acknowledges that the exploration of historical landform genesis is complex and beyond the scope of the manuscript, indicating a limitation in fully understanding the historical genetic forces that may have influenced the current terrain. | The paper compares the proposed graph- driven method with image-driven methods, including four ML methods and one DL method, providing insights into the effectiveness of different approaches for landform recognition. | Accuracy: 91.6% |
| Ewing., et al., **[23]** | The paper was to investigatean alternative approach to terrain characterization Using and remote sensing data and machine learning algorithms. The aimed to predict soil properties such as soil moisture and terrain strength using thermal, multispectral, hyperspectral remote sensing data. | The study of this paper used thermal emissivity from the soils in the morning and afternoon, which limited the | The advantages of provides an alternative approach to terrain characterization using remote approach to an offline application. An online application for soil strength predictions would be desirable for practical use. sensing data, which is more rapid, cost- efficient, and Safer compared to traditional measurements. | Accuracy |

# 3. Methodology

## 3.1 Dataset

The dataset, comprising 10,000 images and occupying 2.56GB of memory, offers a comprehensive collection of satellite-captured remote sensing imagery. It is meticulously categorized into four distinct terrain classes: Sandy, Rocky, Grassy, and Marshy. Each class represents unique land cover characteristics, providing valuable data for various applications such as environmental monitoring, land use planning, and geographical studies. The Sandy terrain images highlight regions dominated by sand, such as deserts and beaches, while the Rocky terrain images depict rugged, stone-covered landscapes. Grassy terrain images capture areas rich in vegetation, including fields and meadows, essential for agricultural analysis and ecological research. Marshy terrain images focus on wetlands, crucial for studying aquatic ecosystems and biodiversity. These high-resolution images, captured from a satellite perspective, facilitate detailed analysis and modeling of Earth's surface. This dataset is instrumental for researchers and professionals in remote sensing, geography, and environmental science, enabling advanced terrain classification and land cover mapping.


***Fig. 1.*** *Four distinct terrain classes*

## 3.2 Existing Methodology

The existing methodology is obtained from the manuscript [21].

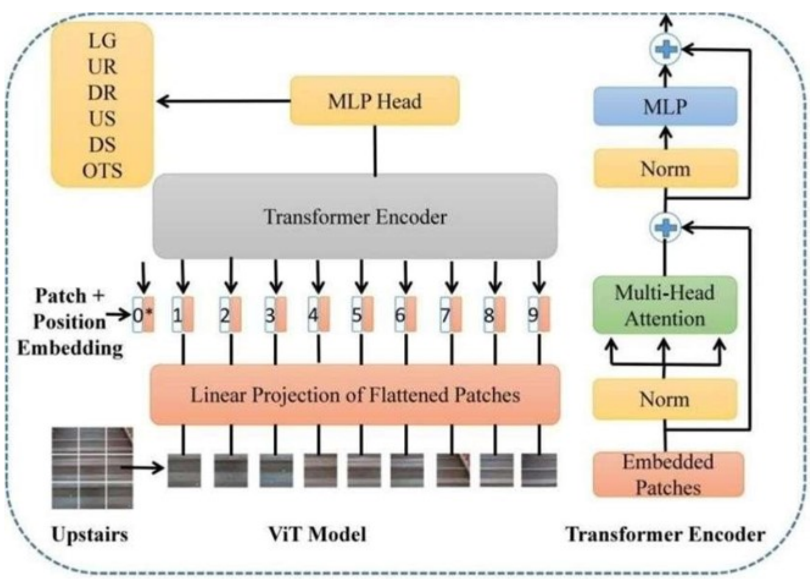



***Fig. 2.*** *Framework of the Existing methodology*

1. **Image Patch Embedding**
   - **Image to Patches**: The input image is divided into a grid of fixed-size patches. For example, an image of size 224×224224×224 might be divided into patches of size 16×1616×16, resulting in 14×14=19614×14=196 patches.
   - **Flattening Patches:** Each patch is then flattened into a 1D vector. For instance, a 16×16×316×16×3 patch is flattened into a 16×16×3=76816×16×3=768 dimensional vector.

2. **Linear Projection of Patches**
   - **Linear Embedding:** Each flattened patch vector is linearly transformed into a fixed-size embedding vector using a learnable linear projection. If the embedding dimension is $D$D, each patch vector of size 768768 is projected to a $D$D-dimensional vector.
   - Patch Embeddings: The result is a sequence of patch embeddings, each of size $D$D.

3. **Adding Positional Embeddings**
   - **Positional Encodings:** Since Transformers do not inherently capture positional information, learnable positional embeddings are added to the patch embeddings. These positional embeddings encode the position of each patch in the original image, helping the model understand spatial relationships.

4. **Transformer Encoder**
   - **Multi-Head Self-Attention:** The core of the Transformer encoder is the multi-head self-attention mechanism. It allows the model to weigh the importance of different patches relative to each other, capturing global dependencies.
   - **Query, Key, Value (Q, K, V):** For each patch embedding, the model computes query (Q), key (K), and value (V) vectors.
   - **Attention Scores:** Attention scores are computed as the dot product of the query and key vectors, scaled by the square root of the embedding dimension, and normalized using a softmax function.
   - **Weighted Sum:** Each value vector is weighted by the corresponding attention score, and the results are summed to produce the output of the attention mechanism.
   - **Multiple Heads:** This process is repeated in parallel for multiple attention heads, allowing the model to capture different types of relationships. The outputs of all heads are concatenated and projected back to the original embedding dimension.

5. **Feed-Forward Network (FFN):** Following the self-attention layer, a position- wise feed-forward network (FFN) is applied, consisting of two linear transformations with a ReLU activation in between.
   - **Layer Normalization:** Each sub-layer (self-attention and FFN) is followed by layer normalization and residual connections, which help stabilize training.

6. **Classification Token (CLS Token)**
   - CLS Token: A special learnable embedding, called the classification token (CLS token), is prepended to the sequence of patch embeddings. This token is used to aggregate information from all patches and is ultimately used for classification.

7. **Final Classification**
   - Classification Head: After processing through the Transformer encoder layers, the final output corresponding to the CLS token is fed into a classification head, typically a fully connected layer, to produce the final predictions.

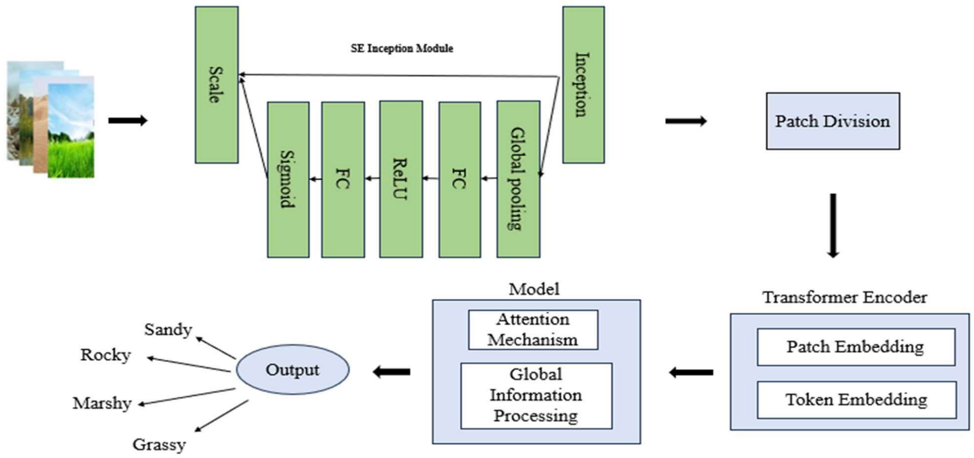



***Fig. 3.*** *Proposed model*

## 3.3 Proposed Methodology

- **Patch Division**
  - The input image is split up into non-overlapping, fixed-size patches. The patch is embedded to obtain tokens that act as input features for subsequent Transformer models.

- **Token Embedding**
  - The patch is a high-dimensional vector in the space. So, they are reduced to tokens that are linear. These tokens are in the format that a transformer can handle.

- **Positional Embedding**
  - Positional embeddings are added to the token embeddings to provide information about the spatial arrangement of the patches of the image. Tokens along with their positional embeddings are the input for the transformer.

- **Transformer Encoder**
  - The transformer encoder consists of multiple layers of self-attention mechanism and a feedforward neural network. This is useful for capturing the relation between the patches enabling us to understand the global context of the image.

- **Attention Mechanism**
  - The embedded patches are processed through an attention mechanism. This allows the model to weigh the importance of different patches to one another, capturing dependencies and context within the image.

- **Global Information Processing**
  - Token representations are generated by the transformer encoder's last layer. These depictions effectively convey the image's overall context. To generate a fixed-size representation of the full image, token representations are frequently combined using pooling operations. These also capture long-range dependencies and understand the realistic context in the image.

- **Sequence and Excitation Block**
  - This module integrates the principles of the Squeeze-and-Excitation (SE) network with an Inception module to enhance the representational power of neural networks. The SE module is designed to recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. This is achieved through two main operations: Squeeze and Excitation.

- **Inception**
  - This part refers to the original Inception module, which is designed to capture multi-scale features by performing convolutions with different kernel sizes and concatenating the outputs. It is adept at extracting diverse spatial features from the input.

- **Global Pooling**
  - This operation aggregates the features across spatial dimensions (height and width), producing a channel-wise summary. The result is a single vector where each element represents a global descriptor of a corresponding channel.
- **First FC Layer**
  - This layer reduces the dimensionality of the pooled vector. It is often followed by a non- linear activation function.
- **ReLu Activation**
  - It is used to introduce non-linearity in the System.
- **Second FC Layer**
  - This layer restores the dimensionality back to the original number of channels.
- **Sigmoid**
  - After the second FC layer, a sigmoid activation function is applied. This outputs values between 0 and 1, effectively creating a set of per-channel scaling factors.
- **Scale**
  - The original input features from the Inception module are multiplied by these scaling factors. Each channel is scaled by the corresponding value from the sigmoid output, effectively reweighting the channel-wise features.
- **Classification Head**
  - This performs task-specific predictions. It transforms the output of the transformer encoder into the class probability of the given image. The output of the transformer encoder creates an aggregated representation of the image that is fed to one or more connected layers with SoftMax activation function for multiclass classification. The final layer produces the class probability.
- **Training**
  - The model is trained using an appropriate loss function, like cross-entropy loss, on a labeled dataset. To minimize the discrepancy between expected class probabilities and actual labels, the model's parameters are modified during training using gradient descent and backpropagation.
- **Fine Tuning**
  - It is performed on the task-specific dataset to adapt the pre-trained model to perform a particular classification task.
- **Testing and Evaluation**
  - After training, the model can be tested on fresh, unseen images to see how well it performs in the task of classifying images. The model is evaluated on metrics such as accuracy, precision, F1 score, confusion matrix, etc.

## 3.4 Working of the Model

- The input image fed into the model is divided into smaller patches. These patches are then fed into the transformer encoder model.
- Each patch is converted into a sequence of tokens (embeddings). These embeddings are numerical representations that capture the essential characteristics of the patch.
- Since the transformer encoder model doesn't naturally know the order of the patches in the original image, positional encoding is added to each token. This positional encoding helps the model understand the relative positions of the patches in the image.
- The transformer encoder is the core of the model. It consists of multiple encoder layers. Each encoder layer uses a self-attention mechanism to attend to the relationships between different tokens. This allows the model to capture long-range dependencies in the image.
- Through the self-attention mechanism, the transformer encoder can capture global information from the entire image, not just individual patches.
- This module refines the feature maps generated by the transformer encoder. It is a type of squeeze- and-excitation network that has been shown to improve the performance of image classification models.

- The input features first go through the Inception module to extract multi-scale features.
- The resulting features undergo Global Pooling to get a summary for each channel.
- This summary is passed through a series of Fully Connected (FC) layers with intermediate ReLU activation, transforming the pooled features.
- A Sigmoid activation generates scaling factors from the transformed features.
- These scaling factors are used to Scale the original features from the Inception module by channel- wise multiplication.
- The output is then sent to the Softmax function to classify them into four classes.

## 4. Results & Discussions

Terrain Recognition has become a crucial need in many applications such as defense, robotics, agriculture, etc. The importance of identifying the nature of the terrain before proceeding with the specific activity has a huge weight. So, terrain recognition has been significant nowadays. There are numerous models for classifying the terrain and recognizing an unseen landform into a particular terrain. Some of the noteworthy models are Vision Transformers, Deep Residual Texture Network, Res NET, etc. However, each of the models is trained in different routines. So, the performance of recognizing an unseen terrain may differ for each model. The performance is calculated in the metrics of Accuracy, Precision, Confusion matrix, etc. Regarding the performance, it’s seen that after performing all the human experiments, Vision Transformers have been proven to provide more satisfactory and precise results in recognizing an unseen terrain.

***Table 2.*** *Performance Analysis for Accuracy, Precision, and Recall*

| Metric | Calculation | Result |
|---|---|---|
| **Accuracy** | (457+510+462+499)/2000 | 0.97 or 97% |
| **Precision (Class 0)** | 457 / (457 + 0) | 1.00 or 100% |
| **Recall (Class 0)** | 457 / (457 + 10) | 0.98 or 98% |
| **Precision (Class 1)** | 510 / (510 + 27) | 0.95 or 95% |
| **Recall (Class 1)** | 510 / (510 + 16) | 0.97 or 97% |
| **Precision (Class 2)** | 462 / (462 + 19) | 0.96 or 96% |
| **Recall (Class 2)** | 462 / (462 + 46) | 0.91 or 91% |
| **Precision (Class 3)** | 499 / (499 + 21) | 0.96 or 96% |
| **Recall (Class 3)** | 499 / (499 + 0) | 1.0 r 100% |

## 5. Advantages and Disadvantages of the Proposed Model

### Advantages

1. Improved Global Information Processing
   - The proposed model uses transformer encoders with self-attention mechanisms, which excel at capturing long-range dependencies and global context in the image. This is a significant improvement over traditional CNNs that primarily focus on local features, leading to more accurate terrain classification.
2. Multi-Scale Feature Extraction
   - The inclusion of an Inception module enables the model to capture multi-scale features, which improves the model's ability to detect details at different scales. This is particularly useful in terrain classification, where the size and shape of terrain features can vary greatly.
3. Channel-Wise Feature Optimization
   - By incorporating a Squeeze-and-Excitation (SE) block, the model can recalibrate feature channels, effectively weighting the important features and reducing noise. This leads to better representational power and improved performance in distinguishing between different terrain types.
4. Efficient Training and Fine-Tuning
   - The model leverages pre-trained architectures such as Vision Transformers, which allows for faster training through fine-tuning. This reduces the computational cost and time needed for model development while improving accuracy on specific tasks like terrain recognition.
5. High Accuracy
   - The proposed model achieves a high accuracy rate of 97%, which indicates its strong ability to generalize across different terrain types. This makes it highly reliable for real-world applications such as navigation, defense, and environmental monitoring.

### Disadvantages

1. Computational Complexity
   - The use of transformers and multi-head self-attention mechanisms increases the computational requirements of the model. It demands high-end hardware, such as GPUs or TPUs, making it less accessible for organizations with limited resources.
2. Large Data Requirements
   - Transformer-based models generally require large datasets to avoid overfitting and achieve optimal performance. While the dataset used (10,000 images) is sufficient, in real-world applications, the model might require more extensive data to generalize to unseen terrains effectively.
3. Potential Overfitting
   - Given the complexity of the model with multiple layers of attention mechanisms and SE blocks, there is a risk of overfitting if not enough regularization techniques, such as dropout or data augmentation, are applied during training.

## 6. Conclusion

Terrain recognition, facilitated by deep learning models, revolutionizes our ability to classify landscapes swiftly and accurately. While various models may exhibit slight differences in classification accuracy, collectively they streamline the task of identifying diverse terrains. This advancement is particularly significant for inaccessible or hazardous areas where human exploration is impractical or risky. By leveraging deep learning algorithms, terrain recognition offers a solution that transcends physical limitations. Rather than sending personnel into potentially dangerous terrain, users can simply upload images, perhaps captured by drones, and input them into the recognition model. The model then autonomously analyzes the images, swiftly providing detailed classifications of the terrain features present. The applications of terrain recognition extend across multiple domains. In defense, it aids in strategic planning by identifying advantageous positions and potential threats. For agriculture, it provides insights into soil types, vegetation cover, and water resources, supporting informed decision-making for crop management and land use. Moreover, terrain recognition contributes to environmental conservation efforts, enabling the monitoring of ecosystems, deforestation, and habitat areas for endangered species. This technology also plays a crucial role in infrastructure development and disaster response. Engineers can utilize terrain data to optimize route planning for roads, railways, or pipelines, while during emergencies, rapid terrain classification guides rescue efforts by identifying safe evacuation routes or areas requiring immediate attention. The integration of terrain recognition with Geographic Information Systems (GIS) enhances its utility, providing decision-makers with comprehensive insights for urban planning, resource management, and disaster risk reduction. Overall, terrain recognition powered by deep learning represents a transformative tool for understanding and navigating our world, offering valuable insights and informed decision-making across various sectors.

## Compliance with Ethical Standards

**Conflicts of Interest:** Authors declared that they have no conflict of interest.

**Human Participants:** The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

## References


[1] Yang, F., Chen, C., Wang, Z., Chen, H., Liu, Y., Li, G., and Wu, X., "Vit- based terrain recognition system for wearable soft exosuit", Biomimetic Intelligence and Robotics, Vol. 3 no. 1, p.100087, 2023.

[2] Moore, J., Stuart, S., McMeekin, P., Walker, R., Nouredanesh, M., Tung, J., and Godfrey, A., "Toward enhanced free-living fall risk assessment: Data mining and deep learning for environment and terrain classification", Intelligence-Based Medicine, Vol. 8, p.100103, 2023.

[3] Li, X., Ma, X., and Song, P., "Fusion of deep feature and hand-crafted features for terrain recognition" In IOP conference series: Materials science and engineering Vol. 646, no. 1, p.012052, 2019.

[4] Iwahashi, J., and Yamazaki, D., "Global polygons for terrain classification divided into uniformslopes and basins" Progress in Earth and Planetary Science, Vol. 9, no. 1, p.33, 2022.

[5] Ramanathan, M., Luo, L., Foo, M. J., Chiam, C. H., Yau, W. Y., & Ang, W. T., "Heuristic vision based terrain recognition for lower limb exoskeletons" In Proceedings of the 16th International Convention on Rehabilitation Engineering and Assistive Technology p. 25-28, 2023.

[6] Li, W., and Hsu, C. Y., "Automated terrain feature identification from remote sensing imagery:a deep learning approach" International Journal of Geographical Information Science, Vol. 34, no. 4, p. 637- 660, 2020.
[7] Kozlowski, P., and Walas, K., "Deep neural networks for terrain recognition task" In 2018Baltic URSI Symposium (URSI) p. 283-286, 2018.
[8] Barrett, A. M., Balme, M. R., Woods, M., Karachalios, S., Petrocelli, D., Joudrier, L., and Sefton- Nash, E., "NOAH- H, a deep-learning, terrain classification system for Mars: Results for the ExoMars Rover candidate landing sites" Icarus, 371, p.114701, 2022.
[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale" arXiv preprint arXiv:2010.11929, 2020
[10] Song, P., Ma, X., Li, X., and Li, Y., "Deep residual texture network for terrain recognition" IEEEAccess, 7, p. 90152- 90161, 2019.
[11] Xue, J., Zhang, H., and Dana, K., "Deep texture manifold for ground terrain recognition" In Proceedings of the IEEEConference on Computer Vision and Pattern Recognition p. 558-567, 2018.
[12] Zhang, H., Xue, J., and Dana, K., "Deep ten: Texture encoding network" In Proceedings of theIEEE conference on computer vision and pattern recognition p. 708-717, 2017.
[13] Li, Z., Zhu, X., Yao, S., Yue, Y., García-Fernández, Á. F., Lim, E. G., and Levers, A., "A largescale Digital Elevation Model super-resolution Transformer" International Journal of Applied Earth Observation and Geoinformation, vol.124, p.103496, 2023.
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I., "Attention is all you need" Advances in neural information processing systems, vol. 30, 2017.
[15] Jiang, Y., Xiong, L., Huang, X., Li, S., and Shen, W., "Super-resolution for terrain modeling using deep learning in high mountain Asia" International Journal of Applied Earth Observation and Geoinformation, Vol. 118, p. 103296, 2023.
[16] Shaban, A., Meng, X., Lee, J., Boots, B., and Fox, D., "Semantic terrain classificationfor off-road autonomous driving" In Conference on Robot Learning p. 619-629, 2022.
[17] Gonzalez, R., and Iagnemma, K., "Deepterramechanics: Terrain classification and slip estimation for ground robots via deep learning" arXiv preprint arXiv:1806.07379, 2018.
[18] Li, Y., Wang, J., Gao, T., Sun, Q., Zhang, L., and Tang, M., "Adoption of machine learning inintelligent terrain classification of Hyperspectral remote sensing images" Computational Intelligenceand Neuroscience, 2020.
[19] Jia, G., Wang, X., and Wei, H., "An effective approach for selection of terrain modeling methods" IEEE Geoscience and Remote Sensing Letters, Vol. 10(4), 875-879, 2013.
[20] Lin, S., Wang, X., Chen, N., and Shen, R., "Directed Positive Negative Terrain Structure Graph Attention Network for Genetic Landform Recognition" IEEE Transactions on Geoscience and Remote Sensing, 2023.
[21] Li, W., Zhou, B., Hsu, C. Y., Li, Y., and Ren, F., "Recognizing terrain features on terrestrial surface using a deep learning model: An example with crater detection" In Proceedings ofthe 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery p. 33-36, 2017.
[22] Chen, Y., Rastogi, C., and Norris, W. R., "A cnn based vision-proprioception fusion method forrobust ugv terrain classification" IEEE Robotics and Automation Letters, vol. 6, no.4, p. 7965-7972, 2021.
[23] Ewing, J., Oommen, T., Thomas, J., Kasaragod, A., Dobson, R., Brooks, C., and Ersal, T., "Terrain Characterization via Machine vs. Deep Learning Using Remote Sensing" Sensors, vol. 23, no. 12, p. 5505, 2023.
[24] Ahmadi, A., Nygaard, T., Kottege, N., Howard, D., and Hudson, N., "Semi- supervised gated recurrent neural networks for robotic terrain classification" IEEE Robotics and Automation Letters, Vol. 6, no. 2, p. 1848-1855, 2021.
[25] Concon, M., Wong, W. K., Juwono, F. H., and Apriono, C., "Deep Learning for Terrain SurfaceClassification: Vibration-based Approach" In ISIC p. 237-243, 2021.