National College of Ireland

TrailSurfNET: Trail Surface Classification Using Convolutional Neural Networks and OpenStreetMap Annotations

MSc Research Project
Msc AI

Mark Finlay
Student ID: x10209221

School of Computing
National College of Ireland

Supervisor: Faithful Chiagoziem ONWUEGBUCHE

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mark Finlay
Student ID:	x10209221
Programme:	Msc AI
Year:	2025
Module:	MSc Research Project
Supervisor:	Faithful Chiagoziem ONWUEGBUCHE
Submission Due Date:	11/08/2025
Project Title:	TrailSurfNET: Trail Surface Classification Using Convolutional Neural Networks and OpenStreetMap Annotations
Word Count:	5,880
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

TrailSurfNET: Trail Surface Classification Using Convolutional Neural Networks and OpenStreetMap Annotations

Mark Finlay
x10209221

Abstract

Accurate classification of hiking trail surfaces is essential for improving outdoor navigation, enhancing safety, and supporting effective land management practices. This research addresses significant data gaps in OpenStreetMap (OSM), where surface-type labels are often incomplete or inconsistent. It introduces and evaluates TrailSurfNET, a novel framework that integrates multi-band Sentinel-2 satellite imagery with OSM annotations to automatically classify trail surfaces.

The study develops a scalable data pipeline that harmonizes diverse OSM tags into five distinct classes (asphalt, paved, gravel, mud/dirt, grass) and generates a balanced dataset through targeted down-sampling. It then conducts a comparative analysis of multiple Convolutional Neural Network (CNN) architectures, including VGG, ResNet-18, ResNet-34, and ResNet-50. The ResNet were evaluated under two conditions: trained from scratch, and fine-tuned with weights pre-trained on the BigEarthNet satellite imagery dataset.

Two key findings emerged: (1) training from scratch matched or exceeded domain-specific transfer learning, and (2) deeper ResNet variants offered a small accuracy gain; the best model (ResNet-50, from scratch) reached 46.73% OA. The findings confirm the viability of using deep learning to augment OSM, providing a scalable solution to enrich critical trail metadata and enhance route planning applications for sustainable trail management.

1 Introduction

1.1 The Data Gap in Modern Trail Navigation

The growing popularity in outdoor recreation, driven by a greater public awareness of its profound physical and mental health benefits (Coventry et al.; 2018; Lackey et al.; 2021), has elevated the importance of high-quality trail systems. This growth creates a corresponding critical need for detailed and reliable trail information. Granular data on trail characteristics, particularly surface type, is paramount for ensuring user safety by allowing for better risk assessment (Cisneros-Frankland et al.; 2023), improving accessibility for a diverse range of users with varying mobility needs (Sanecki et al.; 2023), and enabling more effective and sustainable land management by park authorities (Marion and Wimpey; 2017; Rails-to-Trails Conservancy; 2024). Without such data, the full recreational, health, and economic potential of these natural assets cannot be realized safely or inclusively.

OpenStreetMap (OSM) serves as a vital resource for trail navigation, offering an immediately usable geometry layers for most routes in the UK and Ireland (OpenStreetMap contributors; 2024; Geofabrik GmbH; 2025a,b). As shown in figure 1, the utility of OSM, however, is

hampered by significant gaps in its descriptive data. An analysis of the combined UK and Ireland extract reveals that **only 30% of the 2.79 million ways tagged as paths, footways, or tracks possess a `surface` tag**. Compounding this incompleteness, the crowd sourced tagging vocabulary is often inconsistent, with contributors using ambiguous and overlapping terms like `dirt`, `earth`, or `ground` for similar surfaces. Furthermore, trail surfaces are dynamic and subject to constant change from erosion and environmental factors, meaning even accurately tagged segments can quickly become outdated (Marion and Leung; 2001).

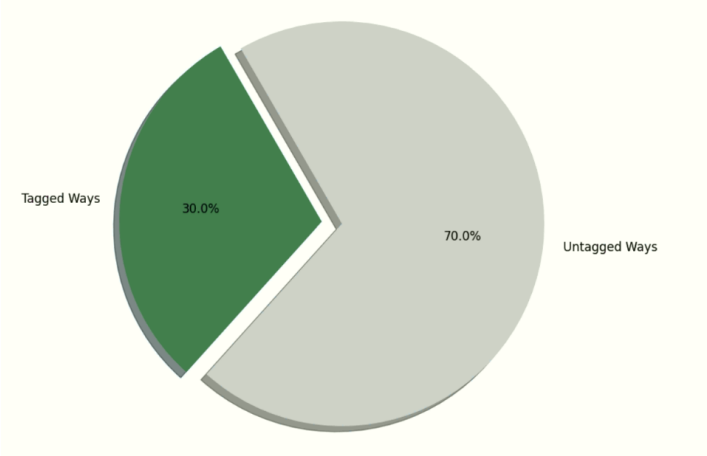


Figure 1: OSM Trail Surface Tag Completeness (UK & Ireland)

surface	n
asphalt	296421
paved	77501
dirt	75690
paving_stones	74839
grass	71989
gravel	52696
ground	50365
unpaved	39513
concrete	35015
compacted	26543
wood	9376
fine_gravel	6727
sett	2656
sand	2335
earth	2277

Figure 2: Top-15 raw `surface=*` values in UK+IE extract.

These persistent gaps in trail-surface information present a clear opportunity for automated data enrichment. The sheer spatial reach of OSM, combined with the noise-tolerant learning capacity of modern Convolutional Neural Networks (CNNs) (Algan and Ulusoy; 2021), creates a timely opportunity to address this challenge. By treating existing OSM tags as provisional—albeit noisy—labels and coupling them with freely available, high-revisit satellite imagery from Sentinel-2, it is now feasible to develop a scalable system for large-scale, up-to-date surface mapping. This approach reframes the data deficiency as a solvable problem, paving the way to automatically enhance OSM with reliable, confidence-rated surface classifications.

1.2 Problem, Aim, and Research Questions

This research addresses the critical challenge of acquiring and maintaining comprehensive and up-to-date surface data for trail networks, which is fundamental for ensuring hiker safety, assessing accessibility, and enabling sustainable land management (Gwerder et al.; 2024; Marion and Leung; 2001). To overcome the deficiencies in existing crowd-sourced data, this research develops and evaluates TrailSurfNET, a scalable deep learning framework. The proposed solution treats existing OpenStreetMap (OSM) surface tags as provisional labels to train a model that can automatically classify trail surfaces from satellite imagery. This approach leverages the inherent robustness of modern deep learning architectures to noise in training data (Algan and Ulusoy; 2021), creating a symbiotic workflow that combines the scale of machine learning with the potential for human-in-the-loop quality assurance.

The primary aim of this research is to develop, evaluate, and validate this scalable framework for the accurate classification of hiking trail surfaces. This is achieved through the following objectives:

- To design and implement a robust data pipeline to systematically collect, filter, and harmonize OSM trail data with multi-band Sentinel-2 imagery, generating a balanced dataset for model training.
- To develop and comparatively evaluate multiple Convolutional Neural Network (CNN) architectures, including a VGG-style network and ResNet models with distinct pre-training strategies.
- To conduct a comprehensive performance analysis of the models using appropriate metrics to assess their effectiveness across the defined surface classes.
- To analyze the broader implications of the results for enhancing the completeness and accuracy of OSM trail data.

This leads to the primary research question: How do convolutional neural network (CNN) models such as VGG-style CNN and ResNet, trained on OSM labeled trail path sections and multi-band Sentinel-2 satellite imagery, perform in terms of accuracy, precision, recall, and F1-scores when classifying trail surface types?

This is supported by the following secondary questions:

- **RQ1:** How accurately can CNN models classify hiking-trail surface types using multi-band Sentinel-2 chips aligned with OSM annotations?
- **RQ2:** What is the impact of model initialization (training from scratch vs. BigEarthNet pre-training) on classification performance for ResNet architectures?
- **RQ3:** How does performance vary across the five consolidated surface classes (asphalt, paved, gravel, mud/dirt, grass)?

1.3 Contribution and Scope

This research makes several contributions to the fields of geospatial machine learning, remote sensing, and volunteered geographic information (VGI). It presents the design and implementation of a scalable data pipeline for integrating OSM vector data with Sentinel-2 imagery to generate labeled image tiles for classification tasks. The work provides a comprehensive comparative analysis of CNN performance, evaluating different model initialization strategies (a VGG-style network from scratch vs. vs ResNet from scratch vs. ResNet with BigEarthNet pre-training) for this specific domain. Finally, it offers an empirical demonstration of the feasibility of using publicly available data to accurately automate trail surface classification, providing a scalable solution to augment and improve critical trail metadata for navigation and land management.

To ensure a focused investigation, the scope of this study is clearly defined. The research concentrates on trail surfaces within the geographic regions of Ireland and the UK, using multi-band Sentinel-2 L2A imagery as the primary data source. The classification is confined to five consolidated surface classes (asphalt, paved, gravel, mud/dirt, and grass) harmonized from raw OSM tags. The methodology is centered on specific CNN architectures and transfer learning strategies, acknowledging that the 10m resolution of Sentinel-2 imagery and the use of a limited temporal range are inherent limitations of this approach.

2 Literature Review

This chapter situates this research within the context of established academic work by reviewing two key domains. First, it examines the state of the art in automated surface classification, focusing on the interplay between remote sensing data and volunteered geographic information (VGI). Second, it synthesizes the literature on the application of deep learning techniques, namely Convolutional Neural Networks (CNNs) and transfer learning, to geospatial vision tasks. This review culminates in the identification of a specific, unaddressed research gap that TrailSurfNET is designed to fill.

2.1 State of the Art in Surface Classification

Prior research into road and trail surface classification reveals a fundamental trade-off between data resolution, coverage, and cost. Studies using high-resolution commercial imagery or piloted aircraft LiDAR achieve high local accuracy but are constrained by high acquisition costs, limited spatial coverage, and infrequent updates (Pešek et al.; 2024; McDermid et al.; 2025). Conversely, analyses using the freely available, global Sentinel-2 dataset benefit from a 5-day revisit cycle and wide area coverage, but must contend with a coarser 10 m resolution, which can increase false-positive rates in detection tasks (Zhou et al.; 2024; Øivind Due Trier and Salberg; 2024). This trade-off is mirrored in the choice of label sources: while manually curated labels provide high-quality ground truth, they are expensive to produce. Volunteered Geographic Information (VGI) from OpenStreetMap (OSM) offers a vast, freely accessible alternative, but its utility is challenged by significant data gaps and inconsistencies.

OSM is a vital resource for geospatial applications, yet its descriptive completeness is a known limitation (Minghini and Frassinelli; 2019). For this study's domain, an analysis of the UK and Ireland reveals that only 30% of the 2.79 million ways designated as paths or tracks carry a `surface=*` tag. Furthermore, the crowdsourced tagging is often ambiguous, with overlapping terms like `dirt`, `earth`, and `ground` used for similar surfaces (OpenStreetMap Wiki contributors; 2023a,b). Despite these issues of noise and incompleteness, a growing body of work demonstrates that the sheer volume of VGI can offset lower label quality, successfully training models for large-scale classification tasks by treating OSM tags as weak labels (Zhou et al.; 2024; Kaiser et al.; 2020).

This research adopts a pragmatic approach by fusing these two powerful, publicly available data sources. We use OSM's noisy but plentiful surface tags as provisional labels for training and Sentinel-2 as the imagery source. This is justified because OSM's geometric accuracy in the UK and Ireland, with a root-mean-square error of approximately 2.3 m, is well-aligned with Sentinel-2's 10 m ground sample distance (Mooney and Corcoran; 2011; Barron et al.; 2014). Furthermore, Sentinel-2's unique spectral richness, including multiple red-edge and Short-Wave Infrared (SWIR) bands, provides critical information for discriminating between vegetation, soil, and man-made materials, compensating for its moderate spatial resolution and making it highly suitable for this classification task (Bill Donatien et al.; 2024).

2.2 Deep Learning for Geospatial Vision

Deep Convolutional Neural Networks (CNNs) are the de facto standard for geospatial image analysis, largely supplanting traditional machine learning pipelines due to their ability to learn hierarchical spatial features directly from pixel data. The choice of architecture, however, depends on the input data's resolution and the training set's volume. Simpler architectures like VGG-style CNN remain effective baselines, achieving high accuracy when data is abundant or imagery is very high resolution (Zhou et al.; 2024; Pešek et al.; 2024). For mid-resolution

satellite imagery like Sentinel-2, deeper architectures with residual connections, such as ResNet, are often preferred (Demir et al.; 2018; Randhawa et al.; 2025). This body of work justifies the comparative evaluation of both VGG and ResNet architectures to determine the optimal model for this specific classification problem.

Beyond architecture, transfer learning is a critical consideration for achieving robust performance. While transfer learning using models pre-trained on large, general-purpose datasets is a common starting point for many computer vision tasks, evidence in the remote sensing field suggests that domain-specific pre-training yields superior results. The advantage of features learned from generic RGB datasets diminishes when models are adapted to the unique spectral characteristics of satellite imagery (Risojević and Stojnić; 2022). Specifically, models pre-trained on large-scale, multi-spectral remote sensing datasets like BigEarthNet (BigEarthNet Project Team; 2025) have been shown to significantly outperform counterparts tuned on general-purpose image datasets. This is because they learn features from the red-edge and Short-Wave Infrared (SWIR) bands that are absent from consumer-grade imagery but are highly informative for discriminating land cover types (Sumbul et al.; 2021). This insight directly motivates this research's experimental design, which explicitly compares the efficacy of training from scratch against a domain-specific, pre-trained approach.

Table 1: Key Studies in Surface Classification Highlighting Methodological Trade-offs

Study	**Sensor & GSD**	**Label Source**	**Key Finding / Limitation**
Zhou et al. (2024)	Google/Maxar HR (0.5 m)	OSM (VGI)	High accuracy is possible with VGI at scale, but results can be skewed by class imbalance.
Pešek et al. (2024)	National Orthophotos (10 cm)	Manual Pixel Labeling	Achieves very high accuracy but requires costly, infrequent aerial surveys and extensive computation.
McDermid et al. (2025)	Piloted Aircraft LiDAR (50 cm)	Manual Trail Tracing	Effective for trail structure but requires bespoke flights and misses trails under dense canopy.
Øivind Due Trier and Salberg (2024)	Sentinel-2 MSI (10 m)	Cadastral Change Detection	Demonstrates national-scale mapping is feasible with Sentinel-2 but suffers from high false-positive rates.

2.3 The Research Gap

The preceding review of sensor technologies, VGI data, and deep learning methods reveals a specific, unaddressed research gap. While remote sensing studies have successfully classified paved and unpaved *road* networks (Zhou et al.; 2024), the fine-grained, multi-class analysis of off-road *trail* surfaces at a national scale remains largely unexplored. This is primarily due to the challenge of sourcing reliable labels, a task made difficult by the sparse and inconsistent

nature of surface attribution in OSM. Furthermore, existing approaches that leverage VGI often lack a principled pipeline for learning from noisy labels (Fonte et al.; 2017; Algan and Ulusoy; 2021) and stop short of proposing an end-to-end workflow that could feed validated, confidence-rated predictions back into the OSM ecosystem (Dalrymple; 2023). This research, therefore, addresses these gaps directly by developing and evaluating TrailSurfNET, a complete pipeline designed specifically to classify multi-class trail surfaces at scale by fusing noisy OSM labels with Sentinel-2 imagery.

3 Methodology

TrailSurfNET couples authoritative vector trail information from OpenStreetMap (OSM) with multispectral Sentinel-2 imagery in order to learn a tile-level mapping from remote-sensing data to five surface classes (*asphalt, paved, gravel, mud/dirt, grass*). At a high level the pipeline is comprised of these main steps:

1. **Data acquisition and spatial preprocessing** that projects, resamples, filters, segments and balances OSM trail geometries, followed by the retrieval of Sentinel-2 L2A tiles as 32×32 pixel images. These images represent **160×160 meter** ground-sampling windows (a **5 m/px** spatial resolution after resampling) centered on every trail segment.

2. **Model Development** in which multiple CNN back-bones are adapted from RGB to nine-channel input (eight spectral bands: B02, B03, B04, B05, B06, B08, B11, B12 + the Sentinel-2 dataMask) and trained under identical data splits.

3. **Quantitative Evaluation Metrics** on a withheld validation set.

3.1 Data Curation Pipeline

3.1.1 OSM Data Filtering, Harmonization and Spatial Projection

The foundation of the training dataset was built from OpenStreetMap (OSM) data, using the latest `.osm.pbf` extracts for Ireland & Northern Ireland and the United Kingdom from Geofabrik (Geofabrik GmbH; 2025a,b). From an initial pool of over 13.8 million line geometries, a focused subset was created by filtering for ways where the `highway` tag was one of `path`, `footway`, or `track`, or the `route` tag was `hiking`. Only ways with a non-null `surface` tag and a geometric length greater than 160 meters were retained to ensure suitability for segmentation.

A key challenge in using VGI is the heterogeneity of crowd-sourced tags. An exploratory analysis revealed 285 unique values for the `surface` tag, many of which were ambiguous or overlapping. These raw tags were consolidated into five distinct and well-represented classes using a deterministic mapping, as detailed in Table 2. Following this, a two-stage balancing process was applied to first the ways and then the resulting segments to counteract class frequency bias, ensuring a balanced dataset for model training.

To guarantee metric accuracy for all subsequent spatial operations, all geometries were re-projected to the EPSG:3035 equal-area coordinate system (EPSG.io; 2025). EPSG:3035 is specifically designed for Pan-European mapping and is the standard grid system for INSPIRE, the EU's spatial data infrastructure framework. While all flat maps of a curved Earth have distortions, ETRS-LAEA (EPSG:3035) is an equal-area projection widely used for pan-European analysis; it preserves area and maintains moderate shape/distance distortions over Europe.

Table 2: Surface Tag Harmonization Scheme

Final Class	Mapped Raw `surface` Tags
`asphalt`	`asphalt, tarmac, macadam, blacktop`
`paved`	`paved, paving_stones, concrete, sett`
`gravel`	`gravel, fine_gravel, pebblestone, compacted`
`mud/dirt`	`dirt, ground, earth, mud, unpaved, sand`
`grass`	`grass, grass_paver`

Table 3: `trail_segments`

Surface Group	Count
Gravel	66551
Mud/Dirt	60460
Grass	55439
Asphalt	36064
Paved	24428

Table 4: `final_balanced_trail_segments`

Surface Group	Count
Asphalt	24428
Grass	24428
Gravel	24428
Mud/Dirt	24428
Paved	24428

3.1.2 Data Balancing and Segmentation

A significant challenge in utilizing OSM data is the inherent class imbalance, where common surfaces are overrepresented. To mitigate the risk of model bias, a two-stage segmentation and balancing strategy was implemented.

First, all harmonized ways from the OSM dataset were programmatically segmented into uniform, non-overlapping 160-meter lengths, as illustrated in Figure 3. The centroids of these new segments were calculated and stored in an intermediate `trail_segments` table. While this step ensured spatial consistency for image tiling, the resulting dataset of 242,942 segments remained highly imbalanced, as shown in Table 3.

To correct this, a final balancing step was applied directly to this set of segments. Each surface class was down-sampled to match the count of the least frequent class, paved. This resulted in the `final_balanced_trail_segments` dataset, containing exactly 24,428 segments per class, as shown in Table 4. Down-sampling was chosen to create a perfectly balanced dataset and avoid the potential for overfitting that can sometimes arise from duplicating minority class samples.

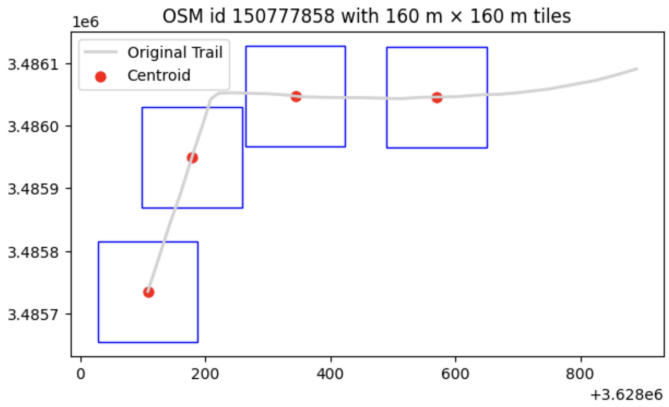


Figure 3: Example of a trail Segment broken into multiple 160x160 m tiles

3.1.3 Sentinel-2 Satellite Data Acquisition

To acquire the necessary satellite imagery for this research, a systematic data collection process was implemented in the notebook `Sentinel 2 Data.ipynb`. The primary data source was the **Sentinel-2 Level-2A (L2A)** collection, which provides atmospherically corrected surface reflectance imagery. Access to this data was facilitated through the **Sentinel Hub API**.

The data acquisition methodology can be broken down into the following key steps:

1. **Geospatial Point of Interest (POI) Identification**: The initial step involved querying a PostgreSQL database to retrieve the central coordinates for each trail segment of interest. These coordinates, stored in a GeoDataFrame, served as the basis for defining the area for which satellite imagery would be requested.

2. **Image Specification using an Evalscript**: A custom `evalscript` was developed to specify the exact data to be downloaded. This script requested eight multispectral bands from the Sentinel-2 L2A product: **Blue (B02), Green (B03), Red (B04), Red Edge (B05, B06), Near Infrared (NIR, B08), and Short-Wave Infrared (SWIR, B11, B12)**. The `dataMask` band was retained as the ninth input channel and used as a feature during training `(0 = invalid/cloud/no-data, 1 = valid)`. Normalization was applied per channel using training-set statistics across all nine channels.

3. **Data Request and Processing**: For each POI, an automated request was sent to the Sentinel Hub API with the following parameters:

    - **Area of Interest**: A **160m x 160m** bounding box was defined around each central coordinate.
    - **Image Dimensions**: The output image was set to a resolution of **32x32 pixels**.
    - **Time Frame**: The query was configured to search for images within the time interval of **January 1, 2024, to December 31, 2024**. To ensure the highest quality data, the `mosaicking_order` was set to "leastCC," which prioritizes images with the least cloud coverage.
    - **Image Resampling**: Bicubic resampling was employed for resampling to maintain image quality.

This automated process ensured a consistent and efficient collection of high-quality, analysis-ready Sentinel-2 imagery for all specified trail segments.

3.2 Model Development and Training

3.2.1 CNN Architectures

Four Convolutional Neural Network (CNN) architectures were selected for a comparative evaluation to identify an optimal approach for classifying trail surfaces from multi-spectral imagery. The chosen models were: a VGG-style network trained from scratch, a lightweight ResNet-18, and deeper ResNet-34 and 50 architectures. This selection allows for a comparison between a standard non-residual architecture and residual networks of varying capacity and pre-training strategies. Key adaptations were made to each model's input layer to accept nine channels, eight spectral bands (B02, B03, B04, B05, B06, B08, B11, B12) plus the `dataMask`, and the final classification layer was modified to output predictions for the **five** target surface classes. The specific configuration and rationale for each architecture are detailed in Table 5.

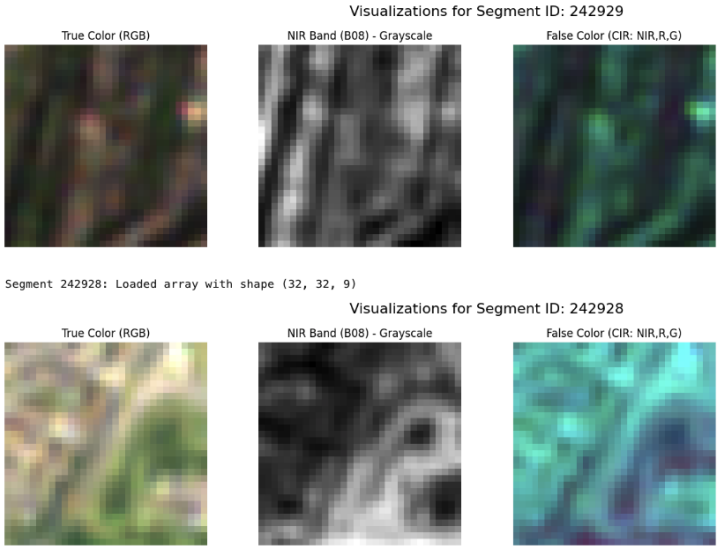


Figure 4: Visualization of Sentinel 2 Data for Segment IDs

Table 5: CNN Backbones and Rationale

Backbone	**Rationale**
VGG-style CNN	Compare against a standard non-residual architecture trained from scratch.
ResNet-18 (from scratch)	Establish a modern residual architecture baseline trained from scratch to isolate the impact of pre-training.
ResNet-18 (BigEarthNet)	Evaluate transfer learning using weights pre-trained on domain-specific Sentinel-2 data on a lightweight residual network.
ResNet-34 (from scratch)	Test if additional model capacity (deeper network) improves performance over a shallower network when trained from scratch.
ResNet-50 (from scratch)	Test if additional model capacity (deeper network) improves performance over a shallower network when trained from scratch.
ResNet-50 (BigEarthNet)	Determine if combining a deeper architecture with domain-specific pre-training yields the highest performance.

3.2.2 Training and Transfer Learning Strategies

While significant efforts were made to mitigate class imbalance at the data level through downsampling, a class-balanced `CrossEntropyLoss` function was used as a complementary, model-level strategy. This weighted loss function serves as a crucial safeguard against any minor imbalances that might persist, ensuring fair representation during training and contributing to a robust and equitably performing model.

Optimization was performed using the `Adam` optimizer with a learning rate of `1e-3` and a weight decay of `1e-4`. To prevent overfitting and ensure convergence, two regularization techniques were employed: a `ReduceLROnPlateau` learning rate scheduler, which reduces the learning rate when validation loss plateaus for 5 epochs, and an `EarlyStopping` mechanism, which halts training if no improvement in validation loss is observed after 12 epochs. All experiments were configured to run for up to 60 epochs with a batch size of 32.

The core of the experimental design was the evaluation of separate initialization and transfer

learning strategies to test their efficacy on this specific geospatial task. The strategies were:

1. **Training from Scratch:** The VGG-style CNN, ResNet-18, ResNet-34 and ResNet-50 models were trained with randomly initialized weights. This strategy serves as a baseline to measure the raw performance of each architecture on the curated dataset and to provide a direct comparison for evaluating the benefit of transfer learning.

2. **Domain-Specific Transfer Learning:** The **ResNet-18 and ResNet-50** models were initialized using weights from a model pre-trained on the `BigEarthNet` dataset (BigEarthNet Project Team; 2025), a large-scale archive of Sentinel-2 satellite images, ResNet-34 was omitted as the pretrained model weights could not be sourced. This strategy was designed to test the hypothesis that pre-training on a large, multi-spectral satellite image dataset is more effective than starting from scratch. For this approach, all layers of the network were unfrozen and fine-tuned on the trail surface data, allowing the model to fully adapt its learned features to the classification task.

3.3 Experimental Evaluation

3.3.1 Quantitative Evaluation Metrics

Reliable assessment of model performance is pivotal for two reasons: (i) to validate that TrailSurfNET genuinely improves on the noisy, volunteer-supplied `surface=*` labels described in the previous section, and (ii) to enable fair comparison with related remote-sensing studies that report a standardized suite of metrics. The main focus of this thesis being *tile-level classification*.

3.3.2 Metrics for Multi-Class Classification

Confusion matrix: All subsequent scores derive from the $C \times C$ confusion matrix $\mathbf{\Phi}$ whose element ϕ_{ij} counts predictions of class j when the reference class is i.

Overall accuracy (OA)"

$$\mathrm{OA} = \frac{\sum_{i=1}^{C} \phi_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} \phi_{ij}}$$

Widely reported but *sensitive to class imbalance*; it can over-state performance when one trail-surface class dominates.

Per-class precision, recall and $\mathbf{F_1}$: For class k

$$P_k = \frac{\phi_{kk}}{\sum_j \phi_{jk}}, \quad R_k = \frac{\phi_{kk}}{\sum_j \phi_{kj}}, \quad F_{1,k} = 2\,\frac{P_k R_k}{P_k + R_k}.$$

We report the *macro-averaged* variants maP, maR, maF1 which weight all surface classes equally and are recommended for imbalanced, multi-class problems.

Retention of Highest Validation Accuracy Model: During each validation epoch, we accumulate predictions and compute overall accuracy. To ensure the optimal model is retained, checkpoints are saved based on the the highest validation accuracy observed throughout the training process.

3.3.3 Implementation for Reproducibility

The entire data curation and modeling pipeline was implemented in `Python 3.11` (Python Software Foundation; 2025), utilizing a suite of open-source libraries to ensure a transparent and reproducible workflow. The core data processing stack was built on `GeoPandas` (GeoPandas Developers; 2025) for handling geospatial vector data, `SQLAlchemy` (Michael Bayer and SQLAlchemy authors; 2025) for database interaction with `PostGIS 3.4` (PostGIS Project Steering Committee; 2025), `PostgreSQL 17` (PostgreSQL Global Development Group; 2025) and `pandas` (The Pandas Development Team; 2025) for tabular data manipulation. Satellite imagery was acquired using the `Sentinel-Hub-Py` library (Sinergise Ltd.; 2024).

For model development and evaluation, the primary framework was `PyTorch 2.3` (PyTorch Foundation; 2025). This was complemented by `scikit-learn` (scikit-learn developers; 2025) for data splitting and `torchmetrics` Lightning AI TorchMetrics Team (2025) for calculating performance metrics. Data visualization, including the generation of confusion matrices and sample image displays, was conducted using `matplotlib` (Matplotlib Developers; 2025).

Model training was performed on Apple M4 Pro, leveraging the MPS backend. To ensure full reproducibility of this research, the complete source code, including the Jupyter Notebooks used for data processing and model training, along with environment configuration files (`requirements.txt`).

The complete source code, including the Jupyter Notebooks and environment configuration files, is hosted in a private GitHub repository and can be made available upon request to ensure full reproducibility of this research.

4 Results

This chapter presents the empirical results from the experiments detailed in the previous chapter. The analysis begins with a comparative evaluation of the six trained models, where we assess overall performance metrics and training dynamics to identify the top-performing architecture. Following this selection, the chapter provides a detailed investigation of this model—the ResNet-50 (scratch). We conduct a quantitative deep-dive into its class-level performance using a confusion matrix and per-class metrics, followed by a qualitative analysis of visual predictions to contextualize its strengths and weaknesses.

4.1 Overall Model Performance Comparison

The primary objective of this experiment was to compare the effectiveness of different CNN architectures and transfer learning strategies. Table 6 summarizes the peak performance of each of the six models on the validation set, evaluated using the metrics defined in Section 3.3.1.

The results show a competitive field of models, with the **ResNet-50 architecture trained from scratch** emerging as the top performer in terms of Overall Accuracy (OA) at 46.73%. Because OA reflects the model's aggregate predictive power across a balanced dataset, it was selected as the primary metric for identifying the best overall model for our subsequent in-depth analysis. It is noteworthy that the **ResNet-18 model pre-trained on BigEarthNet** achieved the highest macro-averaged F1-score (maF1) at 45.85%, though its score was only marginally higher than the ResNet-18(scratch) model (45.81%), suggesting comparable performance in balancing precision and recall.

A key finding from these new results is that **transfer learning did not consistently improve performance**. The models trained from scratch generally performed on par with or

Table 6: Overall Performance Comparison of CNN Models (single run per model; 'ma' indicates macro-averaged)

Model	Pre-trained on	OA	maF1	maP	maR
VGG-style CNN	None (from scratch)	0.44	0.43	0.46	0.44
ResNet-18	None (from scratch)	0.4661	0.4581	**0.4787**	0.4661
ResNet-18	BigEarthNet	0.4644	**0.4585**	0.4741	0.4644
ResNet-34	None (from scratch)	0.4629	0.4568	0.4786	0.4629
ResNet-50	None (from scratch)	**0.4673**	0.4572	0.4782	**0.4673**
ResNet-50	BigEarthNet	0.4614	0.4533	0.4784	0.4614

better than their BigEarthNet-pretrained counterparts. For both the ResNet-18 and ResNet-50 architectures, the models trained from scratch achieved higher overall accuracy than their pre-trained versions. The ResNet-50 (Scratch) performed marginally better (OA: 46.73%) than the ResNet-50 pre-trained on BigEarthNet (OA: 46.14%), a finding that contradicts the expectation that pre-training would provide a performance boost.

When comparing architectures, the **ResNet models universally outperformed the baseline VGG-style CNN**, confirming that residual connections were beneficial for this classification task. The VGG model scored the lowest on all metrics, including an OA of 44.0% and an maF1 of 43.0%. Within the ResNet family, increasing the model depth from ResNet-18 to ResNet-50 seems to benefit the overall accuracy when trained from scratch, with the ResNet-50 achieving the highest OA among all models.

4.2 Class-Level Performance of ResNet-50(Scratch) Model

While overall accuracy provides a high-level view of performance, a deeper analysis of the model's behavior on individual classes reveals important patterns in its predictive capabilities. This section examines the per-class performance of the top-performing ResNet-50 model, trained from scratch, to identify which trail surfaces were classified most successfully and where significant errors occurred.

4.2.1 Correct Predictions

The confusion matrix for the ResNet-50 (Scratch) model (Figure 5) provides a detailed breakdown of correct and incorrect predictions for each of the five surface classes. The center diagonal values represent the number of correctly classified instances. The model was most successful at identifying the **asphalt** and **grass** classes. Specifically, it correctly classified 3,186 instances of asphalt (65.2% of all asphalt samples) and 2,968 instances of grass (60.8% of all grass samples). The **paved** class also showed moderate success, with 1,624 correct predictions, accounting for 33.2% of its true instances. The **mud/dirt** and **gravel** classes proved the most challenging to identify, with correct prediction rates of only 28.8% and 45.6%, respectively.

4.2.2 Sources of Confusion

The off-diagonal values reveal significant sources of confusion between classes. The most prominent error occurred with the **mud/dirt** class, which was most frequently misclassified as **gravel** (1,008 instances) and **grass** (1,739 instances). This highlights a major challenge for the model in distinguishing unpaved surfaces. Similarly, a high degree of confusion was observed

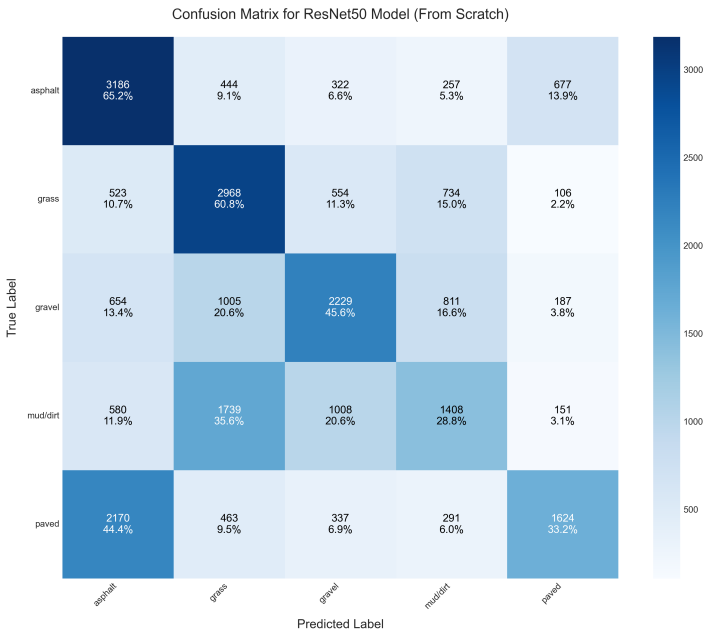


Figure 5: ResNet-50(Scratch) Confusion Matrix

for the **paved** class, which was incorrectly predicted as **asphalt** in 2,170 instances. This is the single largest source of mis-classification in the matrix, indicating that the model struggles to differentiate between a general "paved" surface and one specifically classified as "asphalt." This could be due to visual similarities or overlapping features between these two types of surfaces. Furthermore, the **gravel** class was often confused with **grass** (1,005 instances) and **mud/dirt** (811 instances), suggesting that these unpaved surfaces share similar spectral or textural characteristics that make them difficult for the model to distinguish.

4.3 Per-Class Performance Metrics

The classification report for ResNet-50(Scratch), figure 7, provides quantitative metrics for precision and recall, further clarifying the model's strengths and weaknesses.

Table 7: Classification Report for the ResNet-50 (from scratch) Model

Class	Precision	Recall	F1-Score	Support
Asphalt	0.4479	0.6521	0.5310	4886
Grass	0.4484	0.6076	0.5160	4885
Gravel	0.5009	0.4562	0.4775	4886
Mud/Dirt	0.4022	0.2882	0.3358	4886
Paved	0.5916	0.3324	0.4257	4885
Macro Avg	**0.4782**	**0.4673**	**0.4572**	**24428**
Weighted Avg	**0.4782**	**0.4673**	**0.4572**	**24428**

4.3.1 Precision

The model achieved its highest precision on the **paved** (0.5916) class. This indicates that when the model predicted a surface was paved, it was correct approximately 59.2% of the time. The lowest precision was for the **mud/dirt** class (0.4022), reinforcing that its predictions for this category were the least reliable.

4.3.2 Recall

In terms of recall, which measures the model's ability to find all true instances of a class, asphalt (0.6521) and grass (0.6076) performed best. This means the model successfully identified over 65% of all asphalt instances and over 60% of all grass instances in the validation set. Conversely, the model's most significant weakness was its inability to identify other surface types. The mud/dirt class had a remarkably low recall of only 0.2882, meaning the model failed to find over 71% of the actual mud/dirt surfaces. Similarly, the paved class had the second-lowest recall at 0.3324, indicating a widespread difficulty in correctly classifying these surfaces even when they were present.

4.4 Visualizing Model Predictions

Beyond quantitative metrics, a qualitative evaluation of the model's predictions on sample images provides valuable insight into its real-world behavior. Figure 6 presents a random selection of images from the validation set, along with their true and predicted labels. This visual analysis helps to ground the statistical findings from the confusion matrix and classification report.

The sample predictions confirm the model's strengths and weaknesses. For instance, the model demonstrates a strong ability to correctly classify distinct surfaces like asphalt and grass. However, the visualization also highlights the challenges identified in the error analysis. A notable example is the misclassifications of a mud/dirt surface as grass, visually confirming the confusion between these two spectrally similar, unpaved classes. These examples underscore the difficulty of the classification task and highlight specific areas for potential future model improvement.

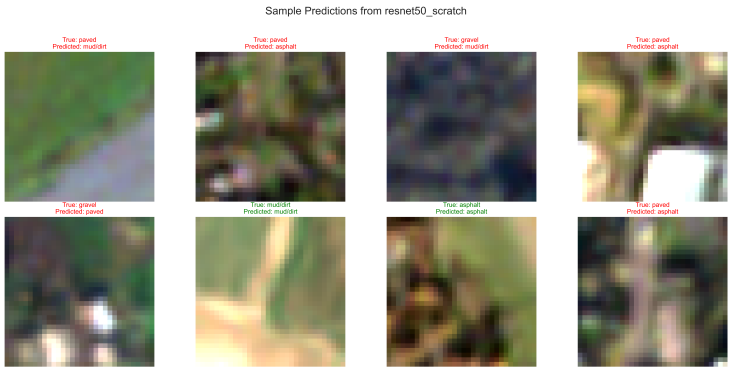


Figure 6: Sample Predictions from ResNet-50 (from scratch) model

5 Discussion

This chapter provides an interpretation of the experimental findings presented in Chapter 4, contextualizing them within the primary objective of classifying trail surfaces using deep learning models on Sentinel-2 satellite imagery. The central finding of this study is the superior performance of the **ResNet-50 (scratch) model**, which achieved the highest overall accuracy of 46.73%.

The results revealed two trends that deserve attention. Firstly, domain-specific transfer learning using BigEarthNet failed to provide a performance benefit; in fact, models trained from scratch consistently performed on par with or better than their pre-trained counterparts. Secondly, within the family of models trained from scratch, increasing architectural complexity

from ResNet-18 up to ResNet-50 yielded a marginal but noticeable improvement in overall accuracy, suggesting that deeper models were better able to capture the intricate features required for this classification task.

Despite these findings, the analysis also confirmed the persistent challenge of distinguishing between spectrally similar unpaved surfaces (e.g., mud/dirt, gravel, grass), which remains the primary factor limiting overall performance. The following sections will delve into the interpretation of these key findings, discuss the study's methodological strengths and limitations, and explore the broader implications of this research.

5.1 Interpretation of Key Findings

5.1.1 The Unexpected Efficacy of Training from Scratch

One of the most significant findings of this study was the consistent and unexpected outperformance of models trained from scratch compared to those using transfer learning with BigEarthNet weights (see table 6). As shown in the overall performance comparison, both the `ResNet-18` and `ResNet-50` models trained from scratch achieved higher overall accuracy than their pre-trained counterparts. This outcome contradicts the initial hypothesis that domain-specific pre-training on a large remote sensing dataset would provide a distinct advantage and warrants a deeper interpretation.

The primary reason for this result may be a **fundamental domain mismatch** between the source (BigEarthNet) and target (TrailSurfNET) tasks, despite both being in the remote sensing field. BigEarthNet is a multi-label land-cover classification dataset where models learn to identify broad categories like `"forest"`, `"pasture"` and `"urban areas"` from Sentinel-2 imagery. The features learned for this task are likely geared towards recognizing large-scale patterns and general land use.

In contrast, the TrailSurfNET task is a fine-grained classification problem that requires distinguishing subtle textural and spectral differences between specific surfaces like `gravel` and `mud/dirt`. The high-level, coarse features from BigEarthNet may not have been sufficiently relevant or adaptable for this more specialized task, making the features learned directly from the trail imagery more effective. Furthermore, the substantial size of the training dataset (over 98,000 images) was likely large enough to allow the models to learn a robust set of specific features from scratch without needing the starting point provided by pre-training.

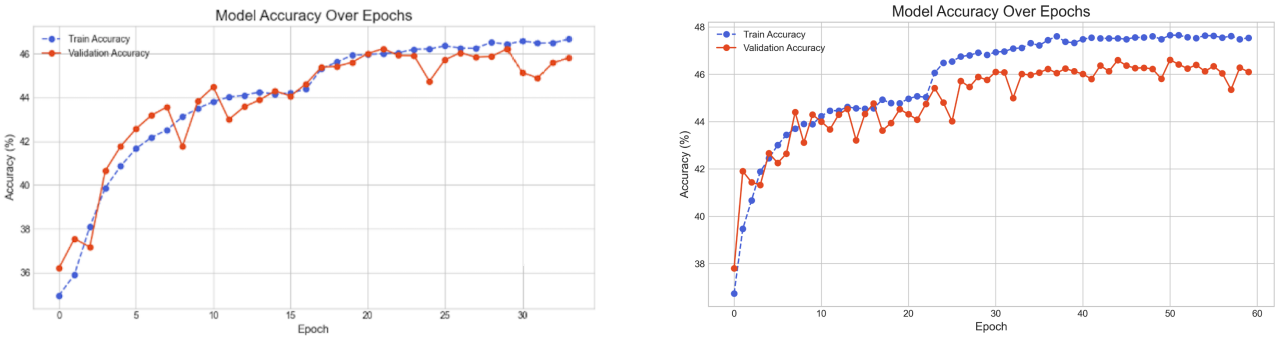
5.1.2 The Impact of Model Complexity and Overfitting

A key finding from the architectural comparison is that **increasing model depth correlated with a marginal but consistent improvement in performance**. The deeper ResNet-50 architecture achieved the highest overall accuracy (46.73%), marginally outperforming its shallower counterparts, ResNet-34 (46.29%) and ResNet-18 (46.61%). This suggests that the greater representational capacity of the deeper network was advantageous for this specific, fine-grained classification task. The additional layers likely enabled the ResNet-50 to learn a more complex hierarchy of features, which was necessary to capture the subtle textural and spectral differences that distinguish the trail surface classes.

However, this performance gain was tempered by the persistent challenge of overfitting. As shown in the accuracy plots (Figure 7), both the ResNet-18 and ResNet-50 models trained from scratch exhibited a clear gap between their training and validation accuracy curves. This indicates that both models began to memorize specifics of the training data rather than learning perfectly generalizable features.

The accuracy curves in Figure 7 illustrate the training dynamics and highlight the challenge of overfitting. Both the ResNet-50 (Figure 7a) and ResNet-18 (Figure 7b) models exhibit a

noticeable gap between their training and validation accuracy curves. This gap, which is a classic indicator of overfitting, appears to widen more steadily in the ResNet-18 model. While the ResNet-50 model also overfit, its validation accuracy consistently remains higher than that of the ResNet-18 throughout the later epochs. This sustained superior performance on unseen data, despite the overfitting, lends further support to its selection as the best-performing model.



(a) ResNet-50(Scratch) Accuracy

(b) ResNet-18(Scratch) Accuracy

Figure 7: Accuracy over epochs for ResNet-50(Scratch) and ResNet-18(Scratch) models.

5.1.3 Analysis of Class Confusion and Data Challenges

A detailed analysis of the ResNet-50 (Scratch) confusion matrix (Figure 5) reveals the specific challenges that limited the model's overall performance. The errors are not random; instead, they point to systematic confusion between spectrally similar classes, a problem exacerbated by data quality issues. The most significant sources of confusion were:

- **Paved vs. Asphalt** The single largest source of error was the model **misclassifying 2,170 paved instances as asphalt**. This indicates that while the model could identify hard, man-made surfaces, it struggled to distinguish between the general paved category and the more specific asphalt class. This suggests significant visual and spectral overlap between these two categories in the training data, making them difficult to separate.
- **Mud/Dirt vs. Grass and Gravel**. The mud/dirt class had the lowest recall (28.8%), being most frequently **misclassified as grass (1,739 instances) and gravel (1,008 instances)**. This highlights a fundamental challenge in separating unpaved, natural surfaces. The confusion with grass likely points to issues of partial vegetative cover or seasonal changes, where muddy paths are partially overgrown. The confusion with gravel suggests that the textural and color features differentiating these two surfaces were not distinct enough for the model to learn reliably.

These errors are compounded by two underlying data challenges inherent to the project's methodology:

1. **Label Noise**: The OpenStreetMap data relies on volunteer mappers who may use tags like `ground`, `earth`, and `dirt` interchangeably, blurring the boundaries between classes and introducing noise into the ground-truth labels.
2. **Resolution Issues**: At a 10-meter resolution, a single Sentinel-2 pixel rarely contains a pure surface. Narrow trails are often captured in "mixed pixels" that also contain surrounding grass, soil, or canopy cover, diluting the spectral signature the model is trying to learn.

5.1.4 Plateau and Breakthrough During Training

A notable dynamic was observed during the training of several ResNet models: an initial performance plateau followed by a distinct "breakthrough" where accuracy would begin to improve again. As illustrated in Figure 7, this pattern was particularly evident in the training histories of the `ResNet-18(Scratch)`, `ResNet-18(BigEarthNet)` `ResNet-34(Scratch)`, `ResNet-50(Scratch)`, and `ResNet-50(BigEarthNet)` models. In these instances, both training and validation accuracy would stagnate for several epochs before resuming an upward trend.

This behavior is likely attributable to the interplay between the optimizer and the learning rate scheduler. The initial plateau suggests that the model had settled into a local minimum or a saddle point in the loss landscape, where small adjustments to the weights at the current learning rate were no longer yielding significant performance gains. The subsequent "breakthrough" often coincides with the activation of the `ReduceLROnPlateau` scheduler, which reduces the learning rate after a period of stagnant validation performance. This reduction allows the optimizer to take smaller, more refined steps, enabling it to escape the local minimum and find a more effective path toward convergence. This recurring pattern demonstrates the critical role of adaptive learning rates in navigating the complex optimization landscapes typical of deep neural networks.

5.2 Methodological Limitations

5.2.1 Spatial Resolution of the Sentinel-2 Data

The 10m spatial resolution of the Sentinel-2 imagery is a fundamental factor that significantly shapes the outcomes of this study. For wider trails and distinct man-made surfaces, this resolution is a powerful asset. It allows the model to reliably differentiate between broad land cover types, such as a paved path contrasted against a grassy park. This capability explains the high performance and accuracy achieved for clear-cut classes like asphalt and grass, where the pixels can capture a relatively pure spectral signature of the surface.

This same 10m resolution becomes a primary constraint when classifying narrower trails or distinguishing between spectrally similar natural surfaces. Many trails are only a few meters wide, meaning they fall within a single pixel that also includes surrounding features like soil, vegetation, or tree cover. This "mixed pixel" effect dilutes the unique signature of the trail surface, making it difficult for the model to perform accurate classification.

5.2.2 Labeling Inaccuracies in the OpenStreetMap Data

A significant methodological challenge stems from the very nature of the OpenStreetMap (OSM) data used for training and validation. As a form of Volunteered Geographic Information (VGI), the labels are crowd-sourced, which makes OSM an incredibly rich but inherently noisy dataset. The quality of the ground truth data is entirely dependent on the diligence and consistency of individual contributors. Labeling errors, subjective interpretations, and outdated information are inevitably present in the dataset. For a supervised learning task, the model's performance is fundamentally tethered to the accuracy of these labels.

This imposes a ceiling on the achievable accuracy of the model, regardless of the sophistication of the network architecture. Even the most powerful model, like the pre-trained ResNet-50, cannot fully overcome the limitations imposed by flawed ground truth data. These labeling inaccuracies introduce noise into the training process, which can explain some of the observed confusion between classes.

5.2.3 Geographic Scope

The geographic scope of this research was intentionally focused on the United Kingdom and Ireland. This decision was driven by several factors, including the relatively high density and quality of OpenStreetMap data in this region, as well as the distinct and well-documented types of trail surfaces commonly found in these countries. By constraining the study area, it was possible to develop a model that is well-attuned to the specific spectral characteristics of the local environment. This focused approach allowed for a more controlled experiment, minimizing the number of confounding variables that would be introduced by a more global dataset.

This limited geographic scope is also a significant limitation. The model, having been trained exclusively on data from the UK and Ireland, is specialized for the environmental conditions and trail construction materials found in this specific part of the world. As a result, its performance is likely to degrade if applied to other geographic regions without re-training or fine-tuning.

5.2.4 Temporal Mis-alignment

An important limitation of this study arises from the temporal mis-alignment between the satellite imagery and the OpenStreetMap (OSM) labels. The analysis was conducted using a composite of Sentinel-2 imagery captured between January and December 2024. However, the corresponding OSM labels were created and last edited by volunteers at various points in time, with many labels potentially pre-dating the image acquisition period by a significant margin. This creates a disconnect where the ground truth label may no longer accurately reflect the real-world surface captured in the imagery.

This temporal discrepancy introduces a form of label noise that is distinct from simple user error. It systematically undermines the model's ability to learn the correct spectral signatures for each surface class, as it is being trained on data where the features and labels are fundamentally out of sync.

5.2.5 Lack of Use of Vision Transformers (ViTs)

This study utilized well-known and proven architectures from the computer vision domain, namely VGG-style CNN and various configurations of ResNet. These models have demonstrated robust performance across a wide array of image recognition tasks and provided a strong baseline for this research. Their standardized nature also ensures that the results are comparable to other studies and that the experimental setup is reproducible.

However, these are general-purpose architectures, designed primarily for processing standard three-channel (RGB) photographic images. They are not inherently optimized to handle the richer information contained within multi-spectral satellite data, such as the 13 spectral bands provided by Sentinel-2. While the models achieved success, it is plausible that more specialized architectures, perhaps those explicitly designed for remote sensing applications or recent advancements like Vision Transformers (ViTs), could have more effectively exploited the unique spectral and spatial patterns in the data.

6 Conclusion

This thesis successfully developed and validated TrailSurfNET, a scalable framework for classifying hiking trail surfaces by fusing public Sentinel-2 satellite imagery with crowd-sourced

OpenStreetMap data. The research demonstrated the viability of using deep learning to address significant data gaps in trail information, a critical step toward improving navigation safety and accessibility.

The investigation confirmed that Convolutional Neural Networks can classify trail surfaces with promising, albeit moderate, accuracy. The best-performing model, a **ResNet-50 trained from scratch**, achieved an overall accuracy of **46.73%**. This headline result was accompanied by two pivotal findings. First, training models from scratch unexpectedly and consistently outperformed domain-specific transfer learning, suggesting the features required for this fine-grained task were best learned directly from the target data. Second, performance varied significantly across classes; while the model excelled at identifying `asphalt` and `grass`, it struggled to differentiate spectrally similar surfaces like `paved`, `mud/dirt`, and `gravel`, highlighting that data quality and resolution are the primary limiting factors.

The journey through this research has yielded not only a functional classification pipeline but also critical insights into the realities of applying machine learning to real-world geospatial problems. It has underscored that even sophisticated algorithms are fundamentally tethered to the quality of the data they learn from. Yet, therein lies the most promising conclusion: the symbiotic relationship between automated systems and human validators. The path forward is not one where machines replace human knowledge, but where they augment it, creating a powerful, virtuous cycle of data improvement. The true potential of this work lies in its ability to serve as a catalyst, focusing the efforts of the global OSM community to create a more accurate, complete, and valuable map for all.

6.1 Future Work

Building on this study, several key avenues for future research are recommended:

- **Data and Sensor Fusion:** Improve accuracy by integrating higher-resolution commercial satellite imagery to mitigate the mixed-pixel problem. Further enhancement could come from fusing imagery with Digital Elevation Models (DEMs) for terrain context or crowd-sourced accelerometer data from smartphones to measure surface roughness.

- **Advanced Model Architectures:** This research deliberately employed established CNN architectures to provide a robust and reproducible baseline. Future work should explore more advanced architectures that could better exploit the rich information in the Sentinel-2 imagery. A particularly promising direction would be the implementation of Vision Transformers (ViTs), which have excelled in other computer vision domains. The self-attention mechanisms in ViTs may be better at capturing the global context and subtle textural patterns within a 160x160 meter tile, potentially offering a performance advantage over the more locally-focused convolutional operations of CNNs.

- **Human-in-the-Loop Systems:** Operationalize the model by developing a human-in-the-loop system. Presenting the model's predictions to the OpenStreetMap community for validation would transform slow manual mapping into an efficient verification task, creating a feedback loop that simultaneously improves OSM data and generates higher-quality labels for re-training.

- **Global Scalability:** Test the adaptability of the TrailSurfNET framework by expanding it to more diverse geographic areas. Re-training and validating the model in different climate zones and continents is a critical step toward developing a more universally applicable tool.

- **Consolidating Classes for a Binary Classification Task:** A highly pragmatic direction for future work would be to simplify the problem from a five-class to a binary classification task. By consolidating the target labels into two broader categories—a **"Hard Surface"** class (comprising `asphalt` and `paved`) and a **"Natural Ground"** class (comprising `grass`, `gravel`, and `mud/dirt`)—the model's utility could be re-evaluated for different applications. This reframing would forgive the model's most common errors, such as the confusion between `paved` and `asphalt`, which are functionally similar for many use cases.

  A projection based on the existing ResNet-50 confusion matrix indicates that this binary model could achieve an overall accuracy of approximately **82.3%**. Such a significant performance increase would make the model immediately more valuable for applications where this simple, high-level distinction is sufficient, such as accessibility routing for wheelchair users or maintenance planning that distinguishes between artificial and natural surfaces. This approach highlights a key trade-off between classification detail and predictive power and offers a path to creating a more robust, application-specific tool.

6.2 Concluding Remarks

This thesis set out with the ambitious goal of addressing a persistent and significant data gap in the world of outdoor recreation: the lack of comprehensive and reliable trail surface information. By developing the TrailSurfNET framework, this research has demonstrated that the fusion of publicly available Sentinel-2 satellite imagery and crowd-sourced OpenStreetMap data offers a viable and scalable solution to this challenge. The investigation confirmed that deep learning models can successfully learn to navigate the inherent noise of VGI to make meaningful classifications, with the ResNet-50 architecture trained from scratch emerging as an effective and efficient model for this fine-grained task.

The journey through this research has yielded not only a functional classification pipeline but also critical insights into the practical realities of applying machine learning to real world geospatial problems. It has underscored the profound impact of data limitations from the spatial resolution of imagery to the temporal mis-alignment and labeling inaccuracies of crowd-sourced data and has highlighted that even the most sophisticated algorithms are fundamentally tethered to the quality of the data they learn from. Yet, therein lies the most promising conclusion: the symbiotic relationship between automated systems and human validators. The path forward is not one where machines replace human knowledge, but where they augment it. The true potential of this work lies in its ability to serve as a catalyst, focusing the efforts of the global OSM community to create a more accurate, complete, and valuable map for all. In doing so, we can make the outdoors safer, more accessible, and more sustainable for generations to come.

References

Algan, G. and Ulusoy, I. (2021). Image classification with deep learning in the presence of noisy labels: A survey, *Knowledge-Based Systems* **215**: 106771.

Barron, C., Neis, P. and Zipf, A. (2014). A comprehensive framework for intrinsic openstreetmap quality analysis, *Transactions in GIS* **18**(6): 877–895.

BigEarthNet Project Team (2025). Bigearthnet: A large-scale sentinel-2 benchmark archive, `https://bigearth.net/#about`. Accessed: 2025-07-06.

Bill Donatien, L. M., Clobite, B. B. and Lemvo, M. M. M. (2024). Comparing sentinel-2 and landsat 9 for land use and land cover mapping assessment in the north of congo republic: a case study in sangha region, *International Journal of Remote Sensing* **45**(22): 8015–8036.

Cisneros-Frankland, D., Gperformance, C. and Sánchez-Gómez, J. (2023). Injuries in trail running: A systematic review, *International Journal of Environmental Research and Public Health* **20**(5): 4499.

Coventry, P., Cooper, C., Gentry, S., Lovell, R., Niemi, M., Rantakokko, M., Varela-Mato, V., White, P. and Garside, R. (2018). Communication with nature: A systematic review of the health benefits of greenspace exposure, *International Journal of Environmental Research and Public Health* **15**(7): 1427.
URL: *https://doi.org/10.3390/ijerph15071427*

Dalrymple, J. (2023). road_surface_classifier, `https://github.com/jdalrym2/road_surface_classifier`.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D. and Raskar, R. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 172–181.

EPSG.io (2025). Epsg:3035 – etrs89 / etrs-laea, `https://epsg.io/3035`. Accessed: 2025-07-06.

Fonte, C. C., Minghini, M., Patriarca, J., Antoniou, V., See, L. and Brovelli, M. A. (2017). Generating and validating land cover maps using volunteered geographic information, *Volunteered Geographic Information and the Future of Geospatial Data*, IGI Global, pp. 51–78.

Geofabrik GmbH (2025a). Download server: Ireland and northern ireland, `https://download.geofabrik.de/europe/ireland-and-northern-ireland.html`. Accessed: 2025-07-06.

Geofabrik GmbH (2025b). Download server: United kingdom, `https://download.geofabrik.de/europe/united-kingdom.html`. Accessed: 2025-07-06.

GeoPandas Developers (2025). About GeoPandas: community-driven geospatial data in Python. Accessed: 2025-08-02.
URL: *https://geopandas.org/en/stable/about.html*

Gwerder, M., Kise, K. and Kawanaka, H. (2024). Accurate hiking time estimation model for mountain trails using sequence-to-sequence learning to improve hiker's safety, *Sensors* **24**(11): 3497.

Kaiser, C., Schultz, M., Zipf, A. and Hahmann, S. (2020). Automatic extraction and filtering of openstreetmap data to generate training datasets for land use land cover classification, *Remote Sensing* **12**(20): 3428.

Lackey, N., Tysor, D., McNay, G., Joy, L., Baker, K. and Hodge, C. (2021). The mental and physical health benefits of a nature-based outdoor activity for people with mental illness: A systematic review and meta-analysis, *PLOS ONE* **16**(9): e0257940.
URL: *https://doi.org/10.1371/journal.pone.0257940*

Lightning AI TorchMetrics Team (2025). TorchMetrics — Machine Learning Metrics for PyTorch and PyTorch Lightning. Accessed: 2025-08-02.
URL: *https://lightning.ai/docs/torchmetrics/stable/*

Marion, J. L. and Leung, Y. F. (2001). Trail degradation as influenced by environmental factors: a state-of-the-knowledge review, *Journal of Soil and Water Conservation* **56**(2): 141–146.

Marion, J. L. and Wimpey, J. (2017). A new generation of science for visitor use management in protected areas, *Parks* **23**(1): 67–78.

Matplotlib Developers (2025). Matplotlib: A 2D Plotting Library for Python. Accessed: 2025-08-02.
URL: *https://matplotlib.org/*

McDermid, G. J., Terenteva, I. and Chan, X. Y. (2025). Mapping trails and tracks in the boreal forest using lidar and convolutional neural networks, *Remote Sensing* **17**(9): 1539. Published 26 April 2025.
URL: *https://doi.org/10.3390/rs17091539*

Michael Bayer and SQLAlchemy authors (2025). SQLAlchemy: Database Toolkit and Object Relational Mapper for Python. Accessed: 2025-08-02.
URL: *https://www.sqlalchemy.org/*

Minghini, M. and Frassinelli, F. (2019). Openstreetmap history for intrinsic quality assessment: Is osm up-to-date?, *Open Geospatial Data, Software and Standards* **4**(1): Add page(s).
URL: *https://doi.org/10.1186/s40965-019-0067-x*

Mooney, P. and Corcoran, P. (2011). Using osm for lbs – an analysis of changes to attributes of spatial objects, *Technical report*, National University of Ireland Maynooth. Available at: `https://mural.maynoothuniversity.ie/id/eprint/5733/`.

OpenStreetMap contributors (2024). About openstreetmap. Accessed: 2025-07-06.
URL: *https://www.openstreetmap.org/about*

OpenStreetMap Wiki contributors (2023a). Tag:surface=dirt - OpenStreetMap Wiki. Accessed: June 22, 2024.
URL: *https://wiki.openstreetmap.org/w/index.php?title=Tag:surface=dirtoldid=2557682*

OpenStreetMap Wiki contributors (2023b). Tag:surface=earth - OpenStreetMap Wiki. Accessed: June 22, 2024.
URL: *https://wiki.openstreetmap.org/w/index.php?title=Tag:surface=eartholdid=2557684*

Pešek, O., Krisztian, L., Landa, M., Metz, M. and Neteler, M. (2024). Convolutional neural networks for road surface classification on aerial imagery, *PeerJ Computer Science* **10**. Published 23 December 2024.
URL: *https://doi.org/10.7717/peerj-cs.2571*

PostGIS Project Steering Committee (2025). Postgis — spatial and geographic objects for postgresql, `https://postgis.net/`. Accessed: 2025-07-06.

PostgreSQL Global Development Group (2025). Postgresql downloads, `https://www.postgresql.org/download/`. Accessed: 2025-07-06.

Python Software Foundation (2025). Download Python – Python.org. Accessed: 2025-08-02.
URL: *https://www.python.org/downloads/*

PyTorch Foundation (2025). PyTorch Foundation — Accelerating open-source AI development. Accessed: 2025-08-02.
URL: *https://pytorch.org/foundation/*

Rails-to-Trails Conservancy (2024). Trail-building toolbox: Maintenance basics, `https://www.railstotrails.org/trail-building-toolbox/maintenance-basics/`. Accessed: June 22, 2025.

Randhawa, S., Aygün, E., Randhawa, G., Herfort, B., Lautenbach, S. and Zipf, A. (2025). Paved or unpaved? a deep learning derived road surface global dataset from mapillary street-view imagery, *ISPRS Journal of Photogrammetry and Remote Sensing* **223**: 362–374.
URL: *https://doi.org/10.1016/j.isprsjprs.2025.02.020*

Risojević, V. and Stojnić, V. (2022). Do we still need imagenet pre-training in remote sensing scene classification? University of Banja Luka, Faculty of Electrical Engineering, Bosnia and Herzegovina.
URL: *https://arxiv.org/abs/2111.03690*

Sanecki, P., Wegrzyn, M., Skowronek, E. and Staszel, A. (2023). Assessing the accessibility of trails in urban forests and parks for people in wheelchairs based on surface and slope parameters, *Sustainability* **15**(10): 7741.

scikit-learn developers (2025). scikit-learn: Machine Learning in Python. Accessed: 2025-08-02.
URL: *https://scikit-learn.org/stable/*

Sinergise Ltd. (2024). Sentinel hub api documentation. Accessed: 2025-07-06.
URL: *https://docs.sentinel-hub.com/api/latest/*

Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B. and Markl, V. (2021). Bigearthnet-mm: A large scale multi-modal multi-label benchmark archive for remote sensing image classification and retrieval, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* .
URL: *https://doi.org/10.1109/MGRS.2021.3089174*

The Pandas Development Team (2025). About pandas — Python Data Analysis Library. Accessed: 2025-08-02.
URL: *https://pandas.pydata.org/about/*

Zhou, Q., Liu, Z. and Huang, Z. (2024). Mapping road surface type of kenya using openstreetmap and high-resolution google satellite imagery, *Scientific Data* **11**(1): 1–1.

Øivind Due Trier and Salberg, A.-B. (2024). National-scale detection of new forest roads in sentinel-2 time series, *Remote Sensing* **16**(21): 3972.
URL: *https://doi.org/10.3390/rs16213972*

AI Usage Disclaimer

This paper was developed with the support of OpenAI's ChatGPT and Google's Gemini. I used these language models extensively throughout the writing process, including rewording sections for clarity and flow, assisting in ideation, organizing the structure of the paper, and interpreting and understanding complex concepts from the literature reviewed. The insights provided by ChatGPT and Gemini were instrumental in enhancing the quality of this paper, although final analysis, interpretations, and conclusions are my own.