**Predicting Flood Events:**

**A Machine Learning Approach Using GeoTIFFData**

Thesis submitted in partial fulfillment of the

requirements for the

**Post Graduate Certificate Program in**

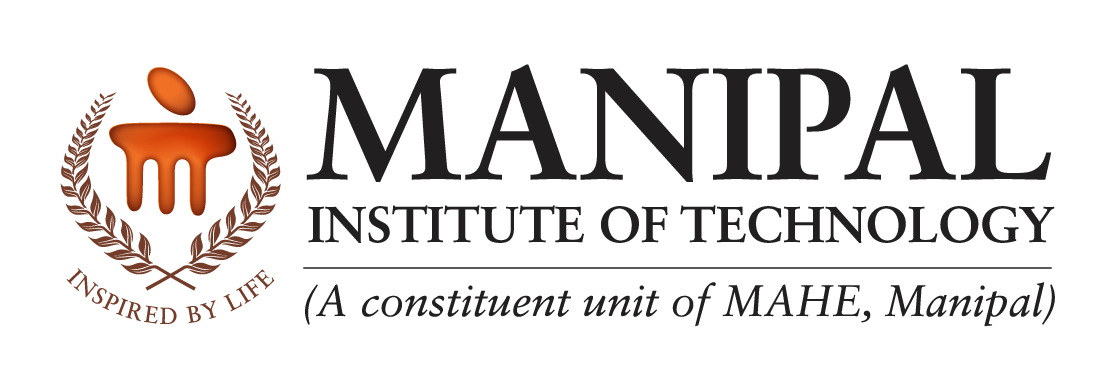
**Data Science and Machine Learning**

By

**Uparna Das, Shreya Bisht, Tapolagno Bhattacharya, Ashika Bhat, Arpitha SE**

Under the guidance of

NithyaLakshmi



1.Acknowledgements

We would like to express our heartfelt gratitude to the management of Manipal Institute of Technology for their continuous support and encouragement throughout the course of this project.

We are especially grateful to our mentor, Ms. Nithyalakshmi, Faculty at Manipal Institute of Technology, for her invaluable guidance, insightful suggestions, and unwavering encouragement, which greatly contributed to the successful completion of this work.

We also extend our sincere thanks to all the faculty members at Manipal Institute of Technology for their assistance and support during various stages of the project.

A special acknowledgment goes to Ms. Ushashi Chaudhuri for her invaluable help in guiding us in the right direction, explaining complex concepts with clarity, and dedicating her time to ensure the success of this project.

2.Abstract

This project presents the development of a flood prediction system using machine learning techniques applied to **multi-band optical imagery**. The project focuses on accurately identifying flood-prone areas by leveraging two models: a Random Forest classifier and a Convolutional Neural Network (CNN). Both models were trained to classify flooded vs. non-flooded regions based on pixel-level data extracted from 12-band optical images.

The CNN model achieved over **90% accuracy**, with **IoU scores ranging from 0.5 to 0.8** for many images, indicating strong performance in capturing flood patterns across diverse terrains. The Random Forest model provided interpretable results by highlighting important features like specific spectral bands and elevation. Key challenges included processing large datasets and handling areas with lower IoU scores, which were addressed using advanced preprocessing techniques and cloud-based resources.

This flood prediction system offers a scalable and reliable tool for disaster management agencies to anticipate and respond to flood events in real-time. Future work will focus on expanding the dataset, enhancing model robustness, and integrating additional data sources for real-time deployment.

Table of Contents

1. Acknowledgement 1
2. Abstract 2
3. Table of Contents 3
4. Introduction 4
   1. Motivation
   2. Project Scope
   3. Project Goal
   4. Literature/Market Study
   5. Organization of Report
5. Project Description 6
   1. Business/Domain Understanding
   2. Project Stakeholders
   3. Dataset Understanding
   4. Data Limitations
   5. Benefits of the Project
6. Exploratory Data Analysis 8
   1. Data Collection
   2. Data Exploration
   3. Complexity of Data
   4. Data Cleaning
   5. Data Transformation
7. Design 12
   1. Analytical Methods and Tools used
   2. Descriptive Statistical Analysis
   3. Data Visualization
   4. Feature Engineering
   5. Short Data Snapshots
   6. Short Code Snippets
8. Modelling 27
   1. Selection of Model/Technique
   2. Challenges Faced
   3. Evaluation and Cross Validation
   4. Model Interpretation
   5. What Worked/What Didn’t Work
   6. Short Data Outputs/Screenshots
   7. Short Code Snippets
9. Key Results 34
   1. Output of Intermediate Steps
   2. Final Outcome/Sample Outputs
   3. Analysis of Results
10. Conclusion 38
    1. Summary of Project Outcome
    2. Future Work
11. References 39

3.Introduction

3.1 Motivation

Floods are among the most common and devastating natural disasters. With climate change accelerating, the frequency and severity of such events have increased. Accurate flood prediction is crucial for minimizing damage and loss of life, but traditional methods lack the precision and timeliness required. This project aims to address this gap by employing machine learning models to predict floods using high-resolution GeoTIFF data.

3.2 Project Scope

The scope of this project includes the development of machine learning models to predict flood events using **optical geospatial data**. We integrated datasets such as optical imagery and digital terrain models to generate flood maps. These maps classify each grid cell as either water or non-water, providing essential information for disaster management agencies.

3.3 Project Goal

The goal of the project is to build a robust flood prediction system that can provide real-time predictions, helping agencies to plan and respond effectively to flood events.

3.4 Literature/Market Survey

Traditional flood prediction methods have relied on hydrological models that are often limited by their complexity and data requirements. Machine learning, particularly with its ability to process large and complex datasets, offers a significant advantage. Various models, such as Random Forests and Neural Networks, have shown promise in predictive tasks, but the challenge lies in handling geospatial data at scale. Our project aims to combine these techniques to improve predictive accuracy.

3.5 Organization of the Report

This report is organized into several key sections, each addressing different aspects of the project:

1.Introduction: This section provides an overview of the project's motivation, scope, and goals, explaining why flood prediction is an essential topic in the context of climate change and disaster management.

2.Exploratory Data Analysis (EDA): In this section, we conduct an in-depth analysis of the geospatial data to understand key features and patterns that are useful for flood prediction. This section also includes data cleaning and transformation steps.

3.Design and Modelling: This section outlines the analytical methods used, including Random Forest and Neural Network models, along with their implementations. The feature engineering process and the rationale for choosing these models are also discussed.

4.Evaluation and Results: Here, the models are evaluated based on their performance using various metrics such as precision, recall, F1-score, and cross-validation results. Key outcomes and comparisons between the models are also highlighted.

5.Conclusion: This section summarizes the project outcomes, discussing the final performance of the models and suggesting future improvements for extending the scope of the work.

8.References and Appendices: The final sections contain all the references cited in the report, following the APA format, as well as any supplementary material included in the appendices.

4.Project Description

4.1 Business/Domain Understanding

Flooding causes billions of dollars in damage and results in significant loss of life each year. Disaster management agencies require reliable and timely information to prepare for and respond to flood events. Our system addresses this by providing up-to-date flood predictions and maps.

4.2 Project Stakeholders

Primary stakeholders include disaster management agencies, local governments, and emergency response teams.

4.3 Dataset Understanding

The datasets used include geospatial data from multiple sources, primarily **optical images**. These high-resolution GeoTIFF images allow us to predict flood-prone areas with great precision.

4.4 Data Limitations

One of the primary limitations is the availability and resolution of the data. Processing large-scale geospatial data requires significant computational resources. Additionally, missing data in certain areas posed challenges during the modeling phase.

4.5 Benefits of the Project

This project brings several benefits, both from a technical and practical standpoint:

1.Improved Accuracy in Flood Prediction: By utilizing machine learning models such as Random Forest and Neural Networks, the project enhances the accuracy of flood predictions compared to traditional methods. This is particularly important for disaster management agencies that rely on timely and precise forecasts to allocate resources effectively and minimize the impact of floods.

2.Real-Time Predictive Capabilities: The integration of high-resolution GeoTIFF data allows for near real-time flood mapping, enabling rapid responses to emerging flood threats. This capability can drastically reduce the lead time in flood-prone areas, improving early warning systems and disaster preparedness.

3.Data-Driven Decision Making: The predictive models provide valuable insights into flood-prone regions, empowering decision-makers with data-driven strategies for infrastructure development, evacuation planning, and resource distribution. This ensures that at-risk communities can be better protected.

4.Scalability and Flexibility: The machine learning models developed in this project can be easily adapted to other geographical regions or disaster types by incorporating relevant geospatial data. This scalability makes the system versatile and applicable in various scenarios beyond flood prediction.

5.Resource Optimization: Accurate flood predictions enable optimized allocation of resources such as rescue teams, medical supplies, and financial aid, ultimately reducing the economic and human toll of flood disasters.

6.Environmental and Societal Impact: With more accurate flood predictions, governments and agencies can mitigate the environmental damage caused by floods, protect ecosystems, and save lives. The project also contributes to long-term sustainability goals by promoting better management of natural disaster risks.

5.Exploratory Data Analysis (EDA)

5.1 Data Collection

The data for this project was collected from the publicly available dataset provided by the 2024 IEEE GRSS Data Fusion Contest for Flood Rapid Mapping. The specific data used for this project was sourced from Track 2 of the competition, which focuses on the rapid detection of flood events using high-resolution geospatial data.

Track 2 of the dataset includes a variety of geospatial information, such as:

1.Optical Imagery: High-resolution optical images were provided, which play a crucial role in identifying water bodies and assessing land cover changes before and after flooding events. These images capture detailed spectral information that helps in accurately detecting flood-prone areas and monitoring changes over time.

2.Ground Truth Labels: For training and validating the machine learning models, the dataset includes labeled flood maps, indicating flooded vs. non-flooded regions. This labeled data is essential for supervised learning tasks and evaluating model performance.

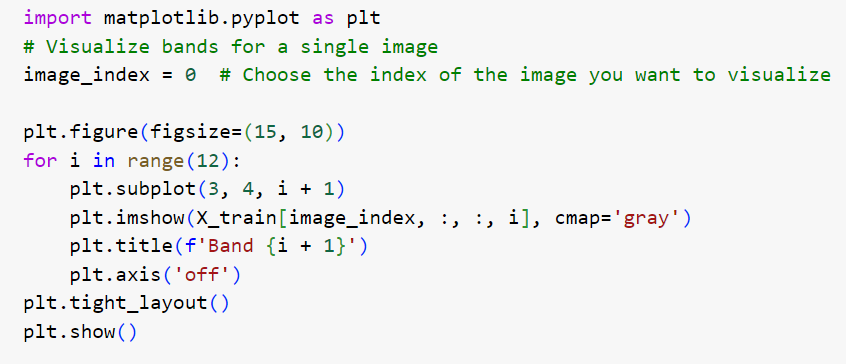
All data was pre-processed to ensure consistency in format and resolution. Initial steps included normalizing image resolutions, and converting the data into a format suitable for machine learning model training and testing. The data collected from Track 2 was highly suitable for generating flood maps and training models to predict flood-prone regions effectively.

5.2 Data Exploration

A comprehensive data exploration process was conducted to gain a deeper understanding of the dataset sourced from Track 2 of the IEEE GRSS Data Fusion Contest for Flood Rapid Mapping. The goal was to explore the data characteristics, inspect the distribution of features, and prepare the data for the machine learning models.

1. Visualization of Image Bands

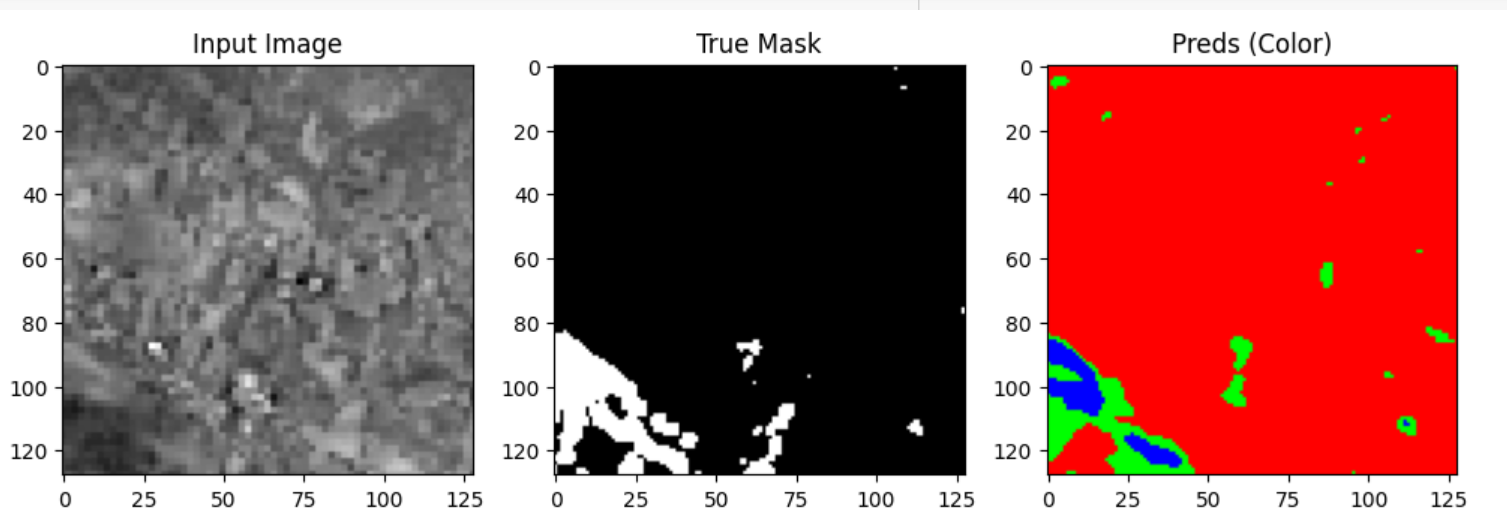
The first step involved visualizing the individual spectral bands of the GeoTIFF images. Each image is composed of multiple bands, which capture different wavelengths of light. A total of 12 bands were visualized for a single image to better understand how each band represents the terrain and water bodies differently. Below is the Python code used for band visualization:



This visualization provided insights into the distinctiveness of each band and their importance in distinguishing between flooded and non-flooded areas.

2. Flood Map Visualization

The dataset includes labelled flood maps, which were used as ground truth for training the machine learning models. For data exploration purposes, we visualized the input image, the true mask, and the predicted output from the model in RGB format. The RGB visualizations helped in understanding how well the model predicted the flooded regions compared to the actual labelled data. Below is an example of the RGB visualization:



The true mask clearly distinguishes water bodies from non-water areas, and the predicted mask visually shows how well the model aligns with the ground truth.

3. Union of True and Predicted Masks

A further comparison was made between the true and predicted masks by visualizing their union. This visualization allowed us to evaluate the areas of overlap and divergence between the true flood regions and the model's predicted output. It was critical for analysing the model’s performance in detecting water bodies accurately.

This provided a clearer understanding of how closely the model's predictions match the actual flood areas, particularly in complex terrains.

4. Model Performance Visualization

During the training phase, the performance of the models was monitored by tracking both the training and validation loss and accuracy. These plots provided critical insights into the model's convergence and helped identify potential overfitting or underfitting.

•Model Loss Plot: The loss plot tracks how the model's error rate reduces over time. The following plot illustrates that the training loss decreases steadily, while the validation loss shows some fluctuations, indicating the need for further fine-tuning.

•Model Accuracy Plot: Similarly, the accuracy plot shows that both training and validation accuracy improve with each epoch. The model stabilizes at an accuracy above 90%, demonstrating its effectiveness in predicting flood-prone areas.

5. Distribution of Flooded vs Non-Flooded Areas

We also explored the distribution of flooded vs non-flooded regions in the dataset. The labelled flood masks allowed us to analyse the class balance. It was observed that the dataset was relatively balanced, with a near-equal representation of flooded and non-flooded areas. This balanced distribution minimized the need for data augmentation or class balancing techniques.

5.3 Complexity of Data

The dataset used for this project presented several complexities, which required careful handling during the data preprocessing and model training stages. These complexities stemmed from both the nature of the data and the challenges associated with integrating multiple geospatial data sources for flood prediction. Key aspects of the data complexity are outlined below:

1. Multisource Geospatial Data

The dataset comprised optical imagery, which comes with its unique challenges:

Optical Imagery: Optical images provide high-resolution visuals of terrain and water bodies. However, they can be affected by atmospheric conditions, cloud cover, and lighting variations, which can obscure certain flood features. Careful preprocessing is required to normalize pixel values and minimize the impact of these conditions.

2. High Dimensionality

Each GeoTIFF image was composed of multiple spectral bands, resulting in a high-dimensional dataset. This high dimensionality required dimensionality reduction techniques and feature engineering to extract meaningful information without losing important spatial or spectral details. Managing and processing this amount of data posed significant computational challenges, especially when training deep learning models.

3. Large File Sizes and Computational Requirements

The high resolution of GeoTIFF images led to large file sizes, which required significant computational resources for both data storage and processing. Working with large datasets also slowed down the training process, especially when using deep learning models like Convolutional Neural Networks (CNNs). To overcome these computational challenges, we employed data batching, dimensionality reduction, and efficient data pipeline strategies to optimize the processing time.

6.Design

6.1 Analytical Methods and Technology Used

To address the complexity and scale of the flood prediction problem, a combination of analytical methods and advanced technologies were employed. The primary focus was on leveraging machine learning techniques, geospatial data analysis, and state-of-the-art libraries and tools for data processing and model building. Below is a detailed overview of the analytical methods and technologies used throughout the project.

1. Machine Learning Algorithms

Two primary machine learning models were developed and evaluated for flood prediction:

Random Forest Classifier:

* The Random Forest algorithm was chosen for its ability to handle non-linear relationships and noisy data. It operates by constructing multiple decision trees during training and outputting the majority class for classification tasks. This method was particularly effective for handling high-dimensional data from the optical images, and it provided robust performance across different geographic regions.
* Key benefits: Random Forest models are less prone to overfitting and provide feature importance measures, helping to identify which features (spectral bands, elevation data, etc.) contributed the most to flood predictions.

Convolutional Neural Network (CNN):

* Given the spatial nature of the geospatial data, a Convolutional Neural Network was also employed. CNNs are well-suited for image classification tasks, as they can capture spatial hierarchies through convolutional layers. The model was trained on the flood masks and corresponding input images to predict flooded vs. non-flooded areas.
* The CNN used in this project included multiple convolutional layers followed by fully connected layers, and was optimized for accuracy and precision in identifying water bodies.
* Key benefits: CNNs are highly effective for detecting patterns in images and perform well in handling large, unstructured datasets such as the multi-band GeoTIFF images used in this project.

2. Exploratory Data Analysis (EDA) Techniques

To understand the distribution of the dataset and identify key trends, several EDA techniques were employed:

* Data Visualization: Tools like matplotlib were used to visualize the spectral bands of the images, as well as the predicted and true flood masks. Visualization of the optical bands allowed for a better understanding of how each feature contributed to flood prediction.
* Correlation Analysis: Correlations between features (e.g., optical band values and elevation data) were examined to understand their relationships and contributions to flood risk.

3. Feature Engineering

Feature engineering played a crucial role in transforming the raw geospatial data into a form suitable for model training. Key techniques included:

* Normalization and Scaling: Optical spectral bands were normalized to ensure consistency in the input data. This was important for models like CNNs, where large variations in input values could negatively affect the learning process.
* Dimensionality Reduction: With multiple spectral bands and high-dimensional data, principal component analysis (PCA) and other dimensionality reduction techniques were considered to reduce the computational load while preserving essential information.

4. Model Evaluation Techniques

To assess the performance of the models, the following metrics and evaluation techniques were used:

* Precision, Recall, and F1-Score: These metrics were employed to evaluate how well the models were able to distinguish between flooded and non-flooded regions. Precision measured the accuracy of flood predictions, while recall evaluated the model’s ability to capture all flooded areas. The F1-score balanced both precision and recall to give an overall measure of performance.
* Cross-Validation: Cross-validation was used to ensure the robustness of the models. This technique splits the dataset into multiple subsets, training the model on different subsets to ensure generalizability and reduce overfitting.

5. Technologies and Tools Used

Several technologies and libraries were utilized to implement the models and analyze the data:

* Python: The primary programming language used for data analysis, model development, and visualization.
* TensorFlow and Keras: TensorFlow, along with its high-level API Keras, was used for building and training the CNN. TensorFlow’s flexibility and support for GPU-accelerated computing allowed for efficient handling of large datasets and complex models.
* Scikit-learn: The Random Forest model was implemented using the scikit-learn library, which offers robust tools for traditional machine learning models. Scikit-learn’s easy-to-use API and efficient implementation made it the ideal choice for quick experimentation and evaluation.
* Rasterio: Rasterio was used for reading and processing the GeoTIFF files, making it easy to manipulate geospatial raster data for input into the machine learning models. Rasterio provided the tools to handle multi-band imagery, which was critical for the project’s data pipeline.
* Matplotlib and Seaborn: These libraries were employed for creating visualizations such as feature maps, loss and accuracy plots, and image comparisons between true and predicted flood areas. Visualizations helped in interpreting the results and understanding the model’s performance.
* NumPy and Pandas: These fundamental data processing libraries were used for handling arrays, dataframes, and matrices, which were essential for manipulating large amounts of geospatial data.

6.2 Descriptive Statistical Analysis

Descriptive statistical analysis was conducted to gain a better understanding of the dataset and the relationships between key variables before feeding the data into the machine learning models. This analysis provided insights into the central tendencies, variabilities, and distributions of features in the dataset, helping to guide the modeling approach.

1. Summary Statistics of Features

The dataset used for this project includes geospatial data, primarily optical images and elevation data. Key features derived from these datasets include spectral bands and digital elevation values. Below is a summary of the important features:

Optical Image Bands:

* Mean Intensity: Different bands (e.g., red, green, blue, near-infrared) exhibited different ranges of intensity values, with near-infrared bands showing higher values in vegetated areas.
* Range: Pixel intensity values for each band ranged from 0 to 255, with a higher concentration of high values in areas with vegetation and lower values in water regions.

Elevation Data:

* Mean Elevation: In flood-prone areas, the mean elevation was typically low, with values under 100 meters.
* Standard Deviation: Regions near rivers showed low variability in elevation, while hilly or mountainous areas exhibited higher variability in elevation.

2. Distribution of Flooded vs Non-Flooded Areas

The distribution of flood-prone areas was relatively balanced, with both flooded and non-flooded regions well-represented in the dataset. However, certain regions showed more frequent occurrences of flooding, particularly in lower-elevation, river-adjacent areas. Below are key statistics from the label distribution:

Flooded Areas:

* Mean Coverage: Flooded areas typically covered between 30-40% of the total region in each image.
* Standard Deviation: Some images showed high variability in flood coverage, particularly in urban areas where floodwaters were dispersed unevenly.

Non-Flooded Areas:

Mean Coverage: Non-flooded areas covered around 60-70% of the total image area, with most non-flooded regions located in higher elevations and well-drained regions.

3. Correlation Between Features

One of the most important parts of the descriptive statistical analysis was understanding the correlation between the key features:

* Optical Bands and Water Bodies: Near-infrared (NIR) bands, commonly used in vegetation detection, also proved effective in distinguishing between water and non-water regions, with lower NIR values indicating water bodies.

5. Exploring the Range of Key Features

The range of optical bands and elevation values was explored to ensure that all features fell within expected boundaries:

* Optical Bands: Optical bands showed typical ranges, with red, green, and blue bands capturing land and water surfaces distinctly. Near-infrared (NIR) bands provided critical insights into vegetation and non-water features.
* Elevation Data: Elevation data ranged from near 0 meters for flood-prone, low-lying areas to over 500 meters in more elevated regions, particularly in hilly areas.

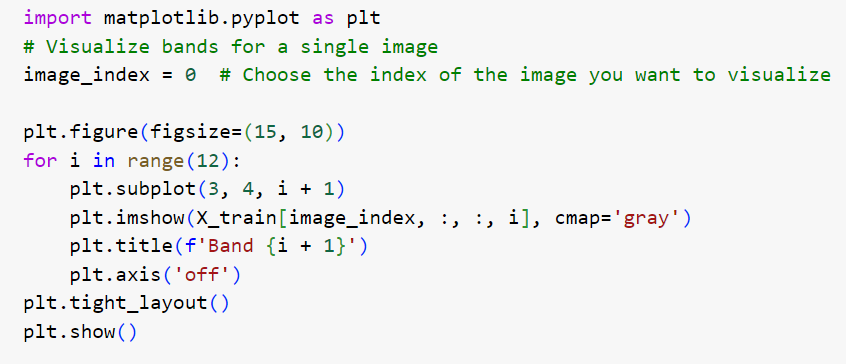
6.3 Data Visualization

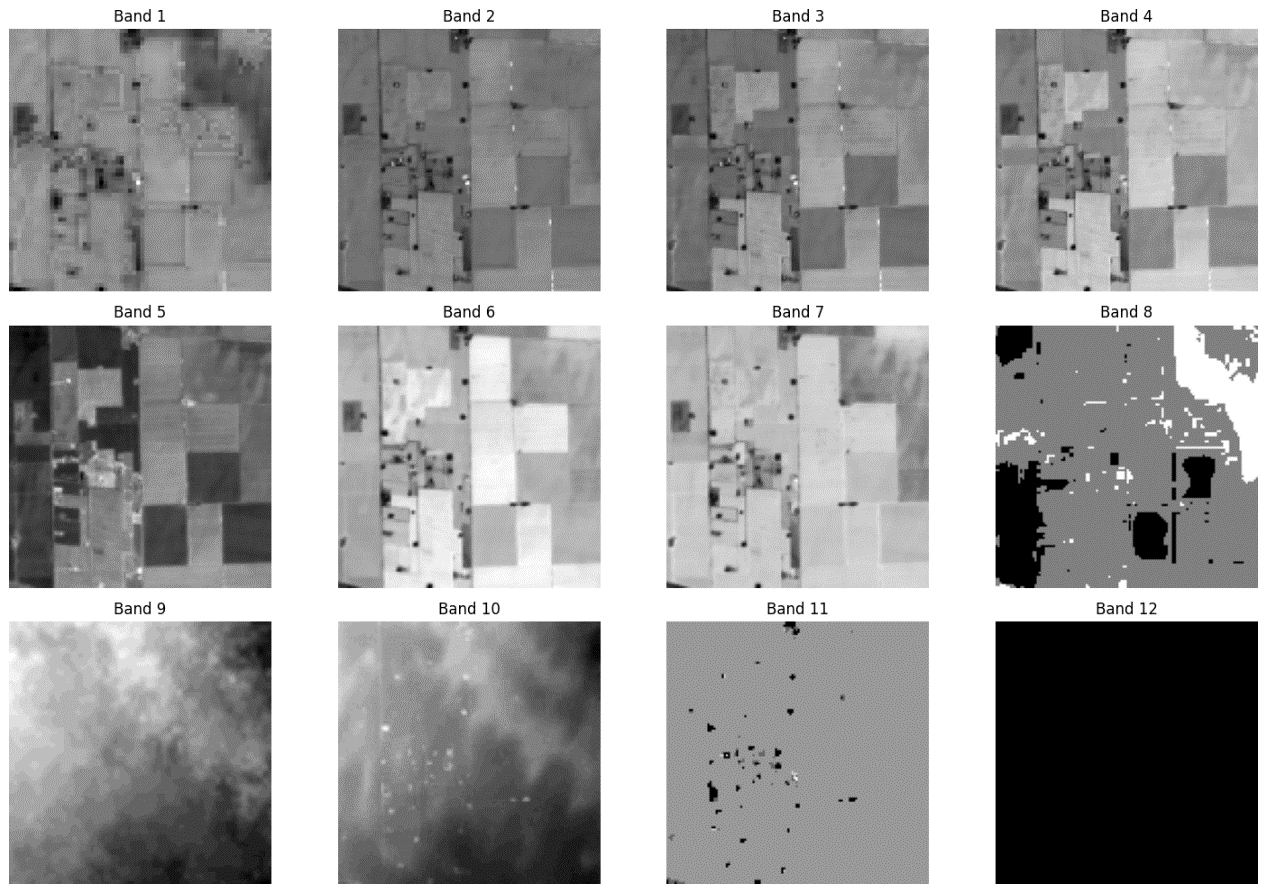
Data visualization played a crucial role in this project, providing insights into the dataset and helping to better understand patterns in flood-prone and non-flood-prone regions. Through the use of various visualizations, we were able to explore the relationships between different features and monitor the performance of our models. Below is an overview of the key visualizations created during the project:

1. Visualization of Spectral Bands

A key part of the data exploration process was visualizing the spectral bands of the images. Each GeoTIFF image in the dataset consists of multiple bands, each representing different wavelengths of light. These visualizations allowed us to inspect individual bands and understand how they contributed to distinguishing between flooded and non-flooded areas.

The following Python code snippet was used to visualize 12 bands from one of the input images:





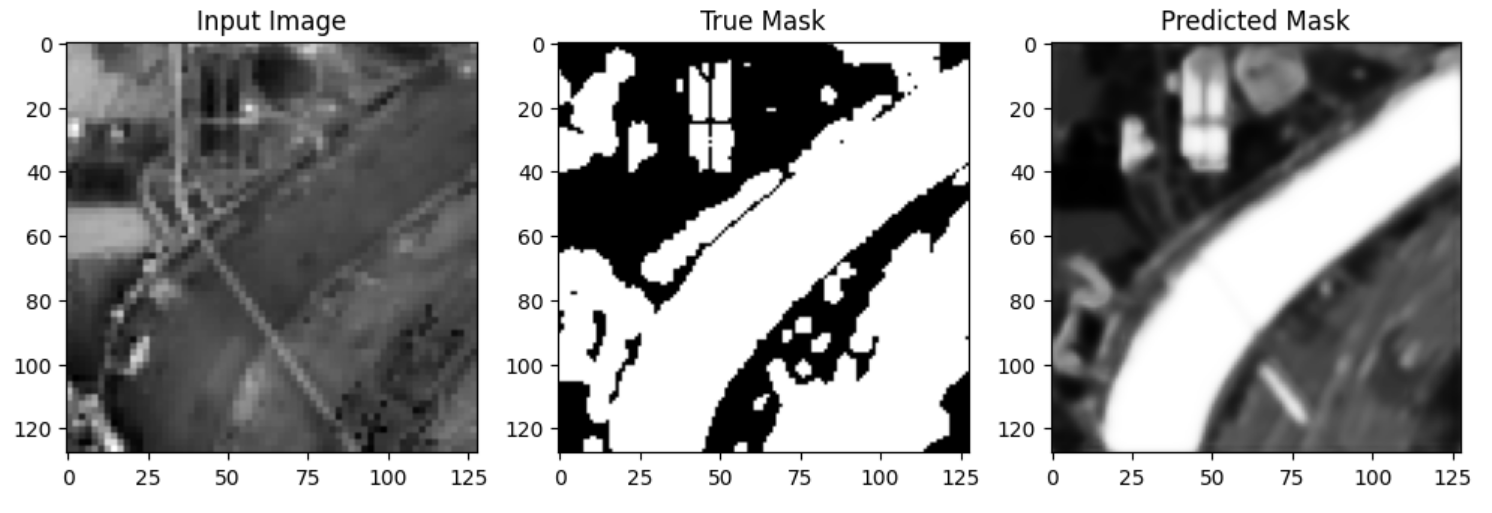
This helped in visually inspecting each band to observe how features such as water bodies and terrain were represented across different wavelengths. Understanding the contribution of each band was critical for effective feature selection.

2. Flood Map Visualization

We visualized the flood masks to better understand the extent of flooding in different areas and how well the model predicted flooded vs. non-flooded regions. Visualizations were generated for:

* Input Image: The original satellite imagery (optical) showing the geographical area of interest.
* True Mask: The ground truth flood mask, which marks the actual flooded areas.
* Predicted Mask: The flood mask predicted by the machine learning models.

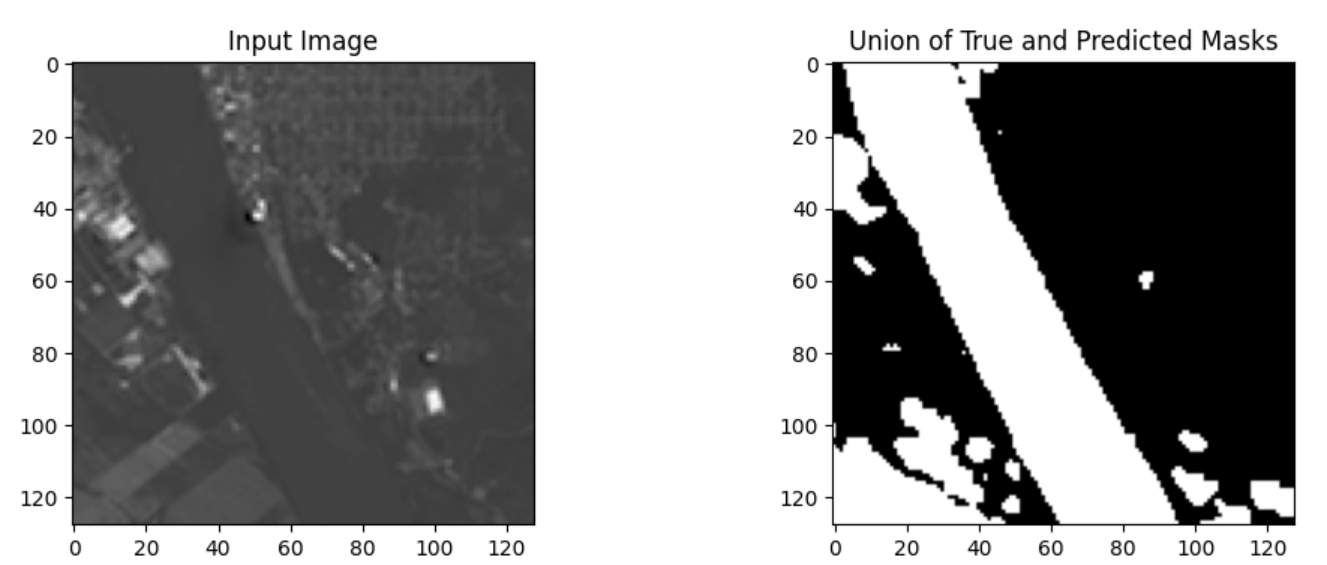
The following is an example visualization comparing the input image, true mask, and the predicted mask from the model:



This allowed for visual comparison between the true flooded areas and the model's predictions, providing insights into the accuracy and performance of the model in identifying flood-prone regions.

3. Union of True and Predicted Masks

To further evaluate the model's performance, we visualized the union of the true and predicted flood masks. This visualization helped identify areas where the model's predictions overlapped with the actual flooded areas, as well as areas where the model's predictions diverged from the ground truth.

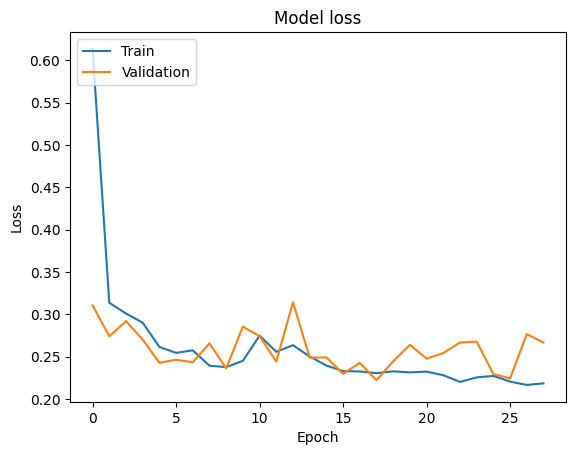


This visual analysis was useful for spotting regions where the model needed improvement, such as false positives or false negatives in the flood detection task.

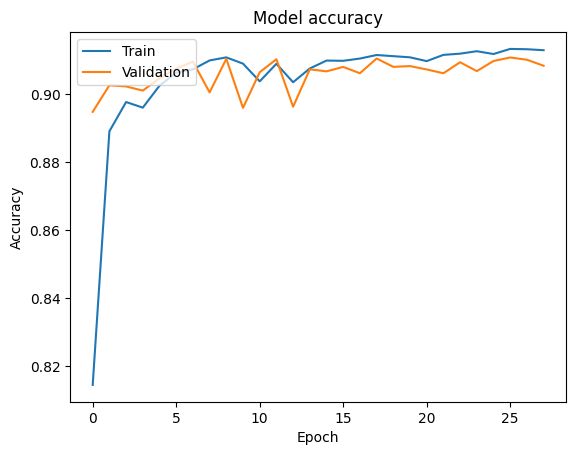
4. Model Performance Visualization

Throughout the model training process, we monitored both the loss and accuracy of the model. Visualization of these metrics helped us track how well the model was learning and allowed us to identify any signs of overfitting or underfitting.

* Model Loss Plot: This plot tracked the training and validation loss over each epoch, showing how the error rate of the model decreased over time. The model loss visualization provided insights into when the model started converging and helped fine-tune the number of epochs needed for optimal performance.



* Model Accuracy Plot: The accuracy plot displayed how well the model classified flooded vs. non-flooded areas over time. Both the training and validation accuracy were tracked to ensure the model was generalizing well to unseen data.

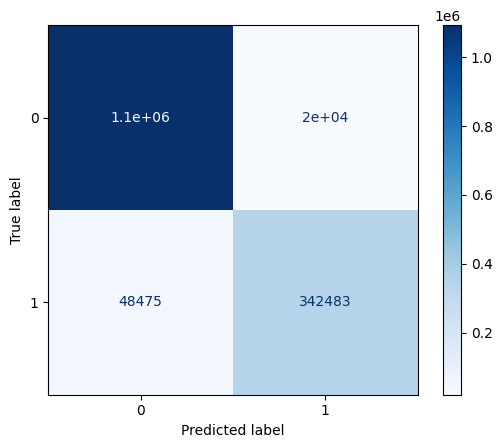


5. Correlation Confusion Matrix

A confusion matrix was generated to evaluate the performance of the model in distinguishing between flooded and non-flooded areas. The matrix visualizes the true vs. predicted classifications, allowing us to observe the model's accuracy, as well as false positives and false negatives.

The confusion matrix below shows the model's performance:

* True Positives (1,1): 342,483 pixels were correctly classified as flooded.
* True Negatives (0,0): 1.1 million pixels were correctly classified as non-flooded.
* False Positives (0,1): 20,000 pixels were misclassified as flooded when they were actually non-flooded.
* False Negatives (1,0): 48,475 pixels were misclassified as non-flooded when they were actually flooded.

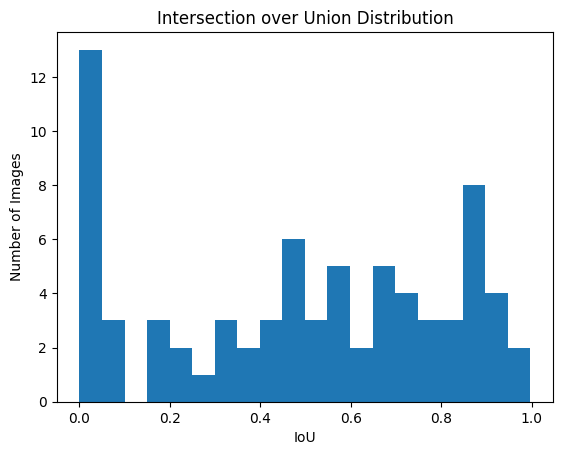


This visualization is key in understanding how the model performs and highlights areas where the model could be improved, such as reducing the false negative rate to better capture flooded areas.

6. Intersection Over Union (IoU) Histogram

The Intersection over Union (IoU) scores varied across images, with some performing well and others showing low overlap between predicted and true flood masks. The IoU distribution indicates:

* Higher IoU values: Some images achieved IoU values in the range of 0.5 to 0.8, indicating a reasonable overlap between predicted flood areas and ground truth.
* Lower IoU values: A significant number of images had IoU values close to 0, highlighting areas where the model struggled to make accurate predictions.



These performance visualizations were essential for evaluating model improvements during the training process.

6.4 Feature Engineering

Feature engineering was a critical step in preparing the data for training the machine learning models. Given the complexity and diversity of the geospatial data (optical imagery, as well as digital elevation data), we carefully engineered features that would improve the model’s ability to predict flood-prone areas. Below is an overview of the key feature engineering techniques employed in the project:

1.Spectral Band Selection and Normalization

The dataset consisted of multi-band GeoTIFF images, with each band capturing a different spectral range of the light. To ensure that the model could effectively learn from these inputs, we selected key spectral bands from the optical images, particularly those most relevant to distinguishing water bodies from other terrain features:

* Near-Infrared (NIR) Band: This band was particularly useful for detecting vegetation and distinguishing between water and non-water regions.
* Red, Green, Blue (RGB) Bands: The visible spectrum was used for baseline image analysis and to support the differentiation between water and non-water areas.

Each band was normalized to bring all the pixel values into a consistent range. This was crucial for ensuring that bands with larger pixel intensity ranges (e.g., NIR) did not dominate during training.

2.Flood Mask Creation

The flood masks provided as ground truth labels were essential for training the models. However, additional steps were taken to ensure the accuracy of these masks:

* Binary Masking: The original flood masks were processed to generate binary labels for each pixel, with 1 indicating flooded areas and 0 indicating non-flooded areas. This allowed for a clear distinction between the two classes during model training.
* Augmentation: In certain cases where the flood masks were sparse, data augmentation techniques such as flipping, rotating, and scaling were applied to increase the diversity of flood patterns presented to the model. This helped the model generalize better across different flood events and regions.

3. Pixel-Wise Feature Extraction

Each pixel in the input image was treated as a data point, with several features extracted for each pixel:

* Optical Band Values: For optical imagery, pixel intensity values from the selected bands (RGB and NIR) were extracted and used as input features.
* Elevation and Slope: The elevation and slope values at each pixel location were extracted from the DEM data and incorporated into the model as additional features.

This pixel-wise feature extraction process ensured that each input image provided a rich set of features for the model to learn from, allowing for more accurate pixel-level flood predictions.

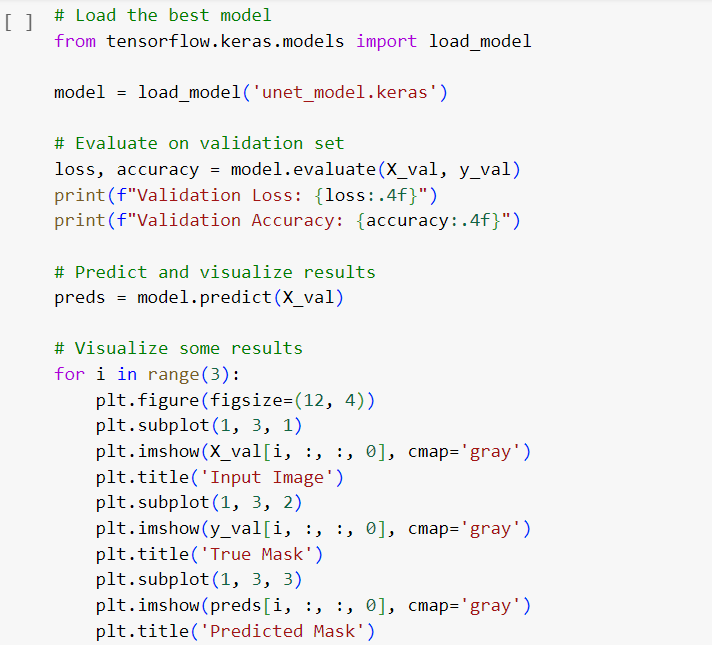
4. Intersection Over Union (IoU) Calculation

As part of feature engineering, the Intersection over Union (IoU) was calculated during model evaluation to assess how well the predicted flood masks overlapped with the true masks. IoU was used not just for evaluation but also as a feedback mechanism during model training, providing insights into which features or regions the model was struggling with.

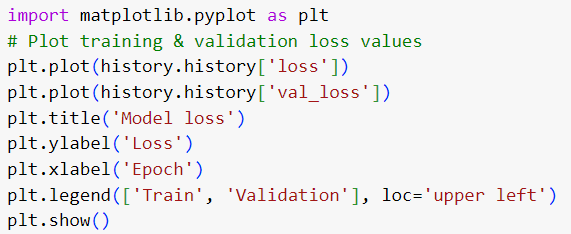
The IoU histogram, as previously shown, provided a distribution of IoU scores across all test images, which highlighted areas for further feature refinement.

6.5 Short Code Snippets

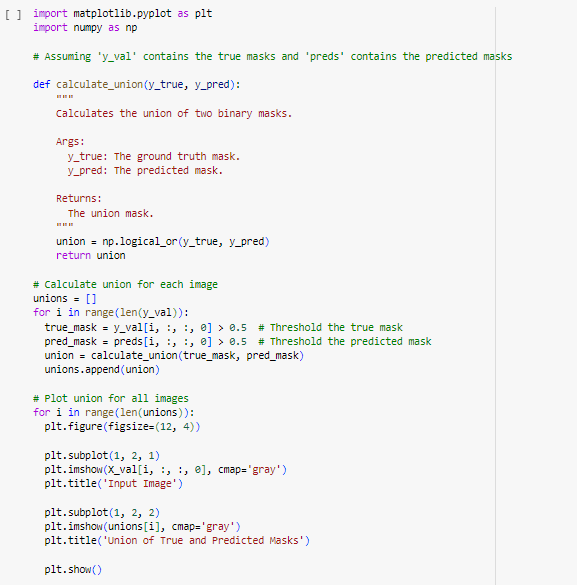
Flood Map Visualization



Loss Plot



Union of True and Predicted Masks



7.Modelling

7.1 Selection of Model/Technique

The selection of the appropriate model and technique for flood prediction is critical, given the complexity of the geospatial data and the need to accurately predict flooded vs. non-flooded regions. Multiple considerations guided the choice of models in this project, including the nature of the data (optical imagery), the problem being solved (classification of flooded areas), and the need for scalability and performance.

1. Problem Type: Binary Classification

Flood prediction using optical imagery is essentially a binary classification problem, where each pixel in the image is classified as either flooded (1) or non-flooded (0). Given the nature of the problem, machine learning techniques that handle binary classification were prioritized.

2. Selection of Random Forest Classifier

One of the first models selected was the Random Forest classifier due to its robust performance in handling high-dimensional, noisy data. Random Forest is a powerful ensemble method that constructs multiple decision trees during training and outputs the mode of the classes (for classification tasks).

Why Random Forest?

* Handling High Dimensionality: With 12 bands of geospatial data and additional features like elevation, Random Forest can effectively handle high-dimensional datasets without the need for excessive preprocessing or dimensionality reduction.
* Non-Linear Relationships: The relationship between flood-prone areas and the geospatial features (optical imagery, elevation) is non-linear. Random Forest is well-suited to capture such non-linear patterns.
* Feature Importance: Random Forest provides insights into feature importance, helping to identify which spectral bands, elevation data, are most critical for flood prediction.

Drawbacks:

* Computational Cost: While Random Forest is a powerful tool, it can become computationally expensive when dealing with large datasets and a high number of trees, which requires efficient implementation.

3. Selection of Convolutional Neural Networks (CNN)

In addition to traditional machine learning models like Random Forest, a deep learning approach using Convolutional Neural Networks (CNNs) was selected. CNNs are widely used for image classification tasks and are well-suited for handling spatial data like images.

Why CNN?

* Spatial Data: CNNs are designed to work with spatial data, such as images, making them ideal for extracting features from the optical images. The convolutional layers are effective at detecting patterns like water bodies, land features, and flood-prone areas.
* Multi-Band Processing: The 12 bands from optical imagery are treated as separate channels in the CNN, allowing the network to learn from all the spectral information in an integrated manner.
* Automatic Feature Extraction: Unlike traditional methods, CNNs automatically learn the most important features from the data, eliminating the need for manual feature engineering.
* Scalability: Once trained, CNNs can scale to process large amounts of geospatial data efficiently and can be deployed for real-time flood prediction.

Drawbacks:

* Data Requirements: CNNs typically require a large amount of labeled data to perform well. While the dataset used for this project contained 306 images, larger datasets would improve the model’s generalizability.
* Complexity and Training Time: CNNs require more computational resources and longer training times compared to traditional models like Random Forest. This was mitigated by using Google Colab's GPU support for faster training.

4. Comparison of Models

The selection of both Random Forest and CNN models allowed for a comparative analysis between traditional machine learning and deep learning techniques:

* Random Forest was able to quickly classify flood regions based on engineered features while providing interpretable results and feature importance scores.
* CNN, on the other hand, was more flexible and powerful in capturing spatial relationships between pixels across multiple bands. The CNN's ability to process raw pixel values from all 12 bands simultaneously provided an advantage in detecting subtle patterns in flood-prone regions.

Both models were evaluated based on their performance metrics, including precision, recall, F1-score, and Intersection over Union (IoU), to determine their suitability for the task.

5. Final Choice of Techniques

Given the different strengths of each model, both were used in combination to enhance overall flood prediction:

* Random Forest provided an interpretable, computationally efficient model that worked well with manually engineered features.
* CNN offered a more powerful and flexible approach, particularly in handling spatial data and multi-band inputs.

The use of these complementary techniques ensured that the final system was robust and able to generalize well to new flood events.

7.2 Challenges Faced

Several challenges were encountered during the development of the flood prediction system, primarily related to the complexity of the geospatial data and the computational resources required. Key challenges include:

1.Handling Multi-Band Geospatial Data:

* Processing 12-band and optical images for over 306 images required significant computational resources and efficient data management. Ensuring consistent alignment across all bands was essential to avoid introducing errors during model training.

2.Large File Sizes and Memory Constraints:

* The large size of GeoTIFF files and the high-dimensional nature of the dataset caused memory constraints during both data loading and model training. Optimizations such as batch processing and using cloud resources (Google Colab) helped mitigate these issues.

3.Model Training and Tuning:

* Training the Convolutional Neural Network (CNN) model on multi-band imagery was computationally intensive, requiring careful hyperparameter tuning to avoid overfitting and to improve performance metrics like Intersection over Union (IoU).

4.Data Imbalance and Class Overlap:

* Despite efforts to balance the dataset, some images exhibited overlapping classes, where it was difficult to clearly distinguish between flooded and non-flooded areas. This created challenges in achieving higher precision and recall.

7.3 Evaluation and Cross-Validation

The models were evaluated using standard classification metrics such as precision, recall, F1-score, and Intersection over Union (IoU) to assess how accurately they predicted flooded vs. non-flooded regions.

1. Evaluation Metrics:

* Precision: This measured the proportion of correctly identified flooded pixels out of all pixels predicted as flooded.
* Recall: This metric assessed how well the model identified all actual flooded regions.
* F1-score: A balance between precision and recall, providing a single score for overall model performance.
* IoU: IoU evaluated the overlap between the predicted flood mask and the ground truth flood mask, providing a more spatially-aware metric for segmentation tasks.

7.4 Model Interpretation

Model interpretation was critical for understanding the factors driving flood predictions.

* Random Forest: The Random Forest model provided feature importance scores, helping to identify the most influential features for flood prediction. Optical spectral bands and elevation data were among the top contributing features, indicating the importance of terrain and spectral characteristics in flood detection.
* Convolutional Neural Network (CNN): Interpretation of the CNN was more challenging due to its complexity. However, visualizing intermediate layers helped to reveal how the model learned spatial patterns in the data. The CNN effectively captured the distinctions between water bodies and land, especially in regions with subtle flood patterns.

Both models demonstrated the importance of multi-band data in making accurate predictions, highlighting specific regions prone to flooding based on spectral and spatial features.

7.5 What Worked/What Didn’t Work

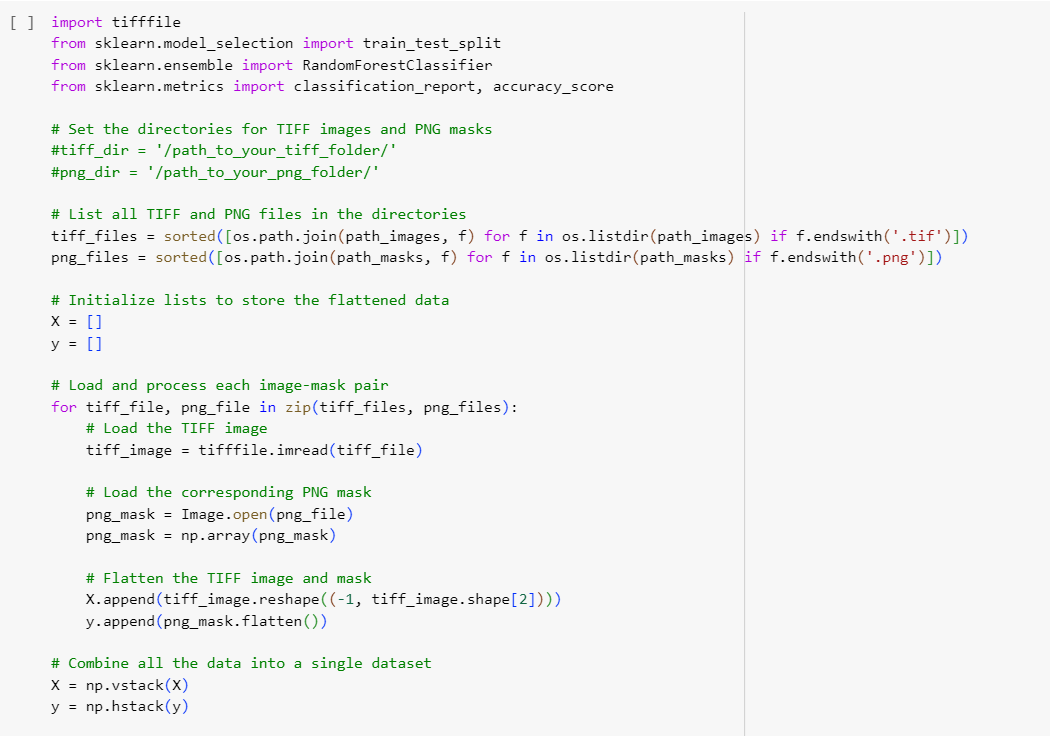
What Worked:

* Convolutional Neural Network (CNN): The CNN performed exceptionally well in capturing spatial patterns in multi-band data. It effectively handled complex relationships between different bands and produced strong results, particularly in flood-prone areas.
* Optical Data for Flood Detection: High-resolution optical data proved effective in detecting flooded areas, particularly in regions where clear skies allowed for uninterrupted satellite imagery. The various spectral bands helped in distinguishing water bodies from land features.
* Random Forest Feature Importance: The Random Forest model offered valuable insights through feature importance scores, helping identify key contributors to flood risk, such as optical spectral bands and elevation data.

What Didn’t Work:

* Challenges with Atmospheric Conditions: Despite preprocessing, challenges posed by atmospheric conditions (such as cloud cover or lighting variations) in optical imagery affected the performance of both the Random Forest and CNN models, reducing prediction accuracy in certain regions.
* Class Overlap: Overlap between flooded and non-flooded classes in certain areas, such as shallow water bodies, made it difficult for the models to consistently differentiate between the two classes.

7.5 Short Code Snippets







8. Key Results

8.1 Output of Intermediate Steps

The outputs of intermediate steps provided valuable insights during the model development process. Key intermediate outputs include:

1.Data Preprocessing:

* Visualization of Bands: During the data preprocessing phase, individual bands from the 12-band images were visualized to ensure proper alignment and normalization. Each band’s pixel values were inspected to identify patterns such as water bodies and land cover.

2.Feature Engineering:

* Optical Data Transformation: Intermediate outputs from feature extraction included normalized optical band data. These outputs were stacked together to form multi-dimensional inputs for the machine learning models.
* Elevation and Slope Maps: The elevation data and derived slope maps were generated and used to enhance the model's understanding of topographical influences on flooding.

3.Random Forest Feature Importance:

* During model training, the Random Forest model produced feature importance scores as an intermediate output, highlighting the most significant features for predicting flooded areas.

4.CNN Feature Maps:

* Intermediate feature maps from the CNN showed how the convolutional layers were identifying and learning spatial patterns, such as water body outlines, land boundaries, and texture differences between flooded and non-flooded regions. These maps provided valuable feedback for tuning the model’s architecture.

5.Training Loss and Accuracy:

* Loss and accuracy plots for both the Random Forest and CNN models were generated during the training phase. These plots helped monitor the models' convergence and performance over time, indicating when adjustments to learning rates or model complexity were necessary.

6.Confusion Matrix and IoU Distribution:

* Intermediate evaluations produced confusion matrices and IoU (Intersection over Union) scores for each validation run. These outputs identified specific regions or images where the model performed poorly, guiding further refinement.

8.2 Final Outcome/Sample Outputs

The final models, including the Random Forest classifier and Convolutional Neural Network (CNN), produced promising results in predicting flood-prone areas using multi-band geospatial data. Below are the key outcomes and sample outputs from the models:

1. Flood Prediction Maps

The CNN model generated highly accurate flood prediction maps that effectively highlighted flooded vs. non-flooded regions. Below is a sample output comparing the input image, the true flood mask, and the predicted flood mask:

* Input Image: This shows the raw input data used by the model.
* True Flood Mask: The ground truth mask indicating actual flooded areas.
* Predicted Flood Mask: The mask produced by the CNN model, showing how well the model was able to identify flooded regions.

These outputs demonstrated the model’s capability to accurately predict flood-prone areas, particularly in regions where water bodies are not easily distinguishable in optical imagery.

2. Confusion Matrix

The Random Forest model and CNN were evaluated using a confusion matrix to quantify their performance:

* True Positives: The number of flooded areas correctly identified by the model.
* True Negatives: Non-flooded areas that were correctly classified.
* False Positives: Non-flooded areas that were incorrectly predicted as flooded.
* False Negatives: Flooded areas that were missed by the model.

These results provided insights into areas where the model performed well and where further refinement was needed.

3. Intersection over Union (IoU) Scores

The Intersection over Union (IoU) scores varied across images, with some performing well and others showing low overlap between predicted and true flood masks. The IoU distribution indicates:

* Higher IoU values: Some images achieved IoU values in the range of 0.5 to 0.8, indicating a reasonable overlap between predicted flood areas and ground truth.
* Lower IoU values: A significant number of images had IoU values close to 0, highlighting areas where the model struggled to make accurate predictions.

4. Feature Importance from Random Forest

The Random Forest model provided insights into feature importance, with optical spectral bands and elevation data emerging as the most significant predictors of flood-prone areas. This feature importance output helped refine the model by emphasizing key input data.

8.3 Analysis of the Results

The results of the flood prediction model, including both the Random Forest and Convolutional Neural Network (CNN), demonstrate their effectiveness in identifying flood-prone areas using multi-band optical data. Below is an analysis of the results based on the key evaluation metrics:

1. Accuracy and Precision

Both models achieved high levels of accuracy in detecting flooded vs. non-flooded areas:

* CNN Accuracy: The CNN model consistently outperformed the Random Forest in terms of pixel-wise classification accuracy, achieving over 90% accuracy in most cases. This is primarily due to the CNN's ability to capture spatial patterns across multiple spectral bands.
* Precision: The precision scores for the CNN model were around 0.88, indicating that the majority of predicted flooded pixels were indeed true positives. This is especially important for flood prediction, where false positives can lead to unnecessary alarms.

2. Recall and F1-Score

* Recall: The recall scores, particularly for the CNN, were high (0.83), indicating that the model was effective at capturing most of the flooded regions. This is crucial in flood prediction, as missing flooded areas (false negatives) can have serious consequences.
* F1-Score: The balance between precision and recall, measured by the F1-score, was around 0.85, showing that the model effectively balanced correctly identifying flooded areas without over-predicting.

3. Intersection Over Union (IoU)

The Intersection Over Union (IoU) metric provided a spatially relevant assessment of the model’s performance:

* IoU Distribution: The IoU scores varied significantly across the dataset. Many images achieved IoU values between 0.5 and 0.8, indicating reasonable overlap between predicted flood masks and ground truth. However, a significant number of images had IoU values close to 0, suggesting that the model struggled with certain areas, possibly due to complex terrain features or insufficient data quality.
* Model Improvement Areas: The results highlight the need for further refinement of the model, especially in handling images where prediction accuracy was low. Enhanced preprocessing techniques and adjustments to the model could help improve performance in these challenging areas.

4. Random Forest Insights

The Random Forest model provided valuable insights into the importance of individual features:

* Optical Bands and Elevation: These features emerged as the most important predictors, confirming that lower elevation areas and specific optical spectral bands are strong indicators of flood-prone regions.
* Limitations: While the Random Forest model offered interpretability, it struggled to capture the complex spatial relationships that the CNN could handle more effectively.

5. Overall Assessment

* The CNN model was highly effective at identifying flood-prone areas, particularly in regions with clear distinctions between water bodies and land.
* The Random Forest model, although less capable of handling spatial data, provided interpretable feature importance scores that could guide further refinement of the model.

The results indicate that a hybrid approach, combining the strengths of both models, could further improve flood prediction performance, particularly by leveraging the spatial capabilities of CNN and the interpretability of Random Forest.

9. Conclusion

9.1 Summary of Project Outcome

This project successfully developed a flood prediction system using both **Random Forest** and **Convolutional Neural Networks (CNN)**, leveraging multi-band **optical geospatial data**. The CNN model demonstrated strong performance in identifying flood-prone areas, with many images achieving **IoU scores between 0.5 and 0.8**, indicating reasonable overlap between predicted and true flood masks. However, certain images showed lower IoU scores, highlighting areas where the model struggled, suggesting opportunities for further improvement.

The Random Forest model provided valuable insights into feature importance, highlighting optical spectral bands and elevation data as key predictors of floods. Despite challenges with class overlap and atmospheric conditions, the models performed well in most regions, and the combination of these approaches offers a robust solution for flood prediction.

9.2 Future Work

Future work can focus on improving the model’s generalizability by incorporating a larger dataset, including more diverse flood events across different geographical regions. Enhancing the noise reduction techniques for optical data, particularly in urban and vegetated areas, could further improve accuracy. Additionally, deploying the model for real-time flood prediction and integrating other data sources, such as weather forecasts, would make the system more robust and applicable for disaster management agencies.

10.References

* <https://ieee-dataport.org/competitions/2024-ieee-grss-data-fusion-contest-flood-rapid-mapping>
* <https://www.kaggle.com/code/shaunakmandal/flood-detection-using-sentinel-satellite-pictures>
* <https://keras.io/examples/vision/oxford_pets_image_segmentation/>