Data preparation: What did you do?

To prepare the dataset for training, I first loaded the dataset by using Keras preprocessing image_dataset_from_directory to load images from the project folder. The folder contained two sub-directories, damaged (images of buildings with damage) and undamaged (images of buildings without damage). Then, I split the dataset into 80% training and 20% testing data to train the model and be able to evaluate it. I also printed out the number of images and how many are in the training and testing data to ensure the dataset is correct. Next, I resized the images to (128, 128) to ensure that the input dimensions were consistent. Finally, I normalized pixel values to the range [0, 1] by applying a rescaling layer and implemented data augmentation techniques like rotation, zoom, and horizontal flipping to enhance model generalization.

Finally, I printed out a few images with the label to make sure that they were what I was expecting and that there were no random pictures that were not satellite images. I also printed out sample images from each class to understand the data distribution.

Model design: Which architectures did you explore, and what decisions did you make for each?

Fully Connected ANN (Artificial Neural Network):

For the fully connected ANN, it flattens the layer to convert image data into a 1D vector. I used three dense layers, 128 neurons with ReLU activation, 64 neurons with ReLU activation, and 1 neuron with a sigmoid activation for binary classification. I think ANN is a simple approach but doesn't have the ability to extract spatial features, making it less useful for image classification. It does work well on structured/tabular data but performed worse than CNNs for this dataset.

LeNet-5 Convolutional Neural Network (CNN):

The LeNet-5 CNN has two convolutional layers, one with 6 filters (5×5) followed by MaxPooling and one with 16 filters (5×5) followed by MaxPooling. It has three fully connected layers, each with 120 neurons with ReLU activation, 84 neurons with ReLU activation, and 1 neuron with a sigmoid activation for binary classification. I think that LeNet-5 is a classic CNN and performed better than the ANN due to feature extraction. However, it was designed for small grayscale images (28×28), so it may not be good for 128×128 RGB images like the satellite images are.

Alternate LeNet-5 Architecture:

The alternate LeNet-5 CNN followed the provided paper, which has three convolutional layers, one with 32 filters (3×3) followed by MaxPooling, one with 64 filters (3×3) followed by MaxPooling, and two with 128 filters (3×3) each followed by MaxPooling. It has dropout that has hyperparameter tuning to prevent overfitting, 512 neurons with ReLU activation, and 1

neuron with a sigmoid activation for binary classification. I think this CNN performed the best because there were more filters and smaller kernel sizes to capture more detailed patterns in images, dropout helps prevent overfitting, and deeper feature extraction. However, after running the model a few times and making small changes, I found that it was still overfitting just not as severely as the original LeNet-5 CNN. In addition, I also found that having the hyperparameter tuning for the dropout made training the model very slow.

Model Training:

Finally, I trained each model using the training dataset and then validated it using the validation set.

Model evaluation: What model performed the best? How confident are you in your model?

After training and evaluating the three architectures, their accuracy and loss on the validation dataset are below.

Model	Training Accuracy	Validation Accuracy	Notes
Fully Connected ANN	76.98%	71.25%	Struggled because of the lack of spatial feature extraction.
LeNet-5 CNN	99.79%	92.78%	Captured spatial features but slightly overfitted.
Alternate LeNet-5 CNN	98.52%	97%	Best performance, deeper network + dropout reduced overfitting.

I think it performed better because more convolutional layers allowed for deeper feature extraction, and the 0.5 dropout reduced overfitting and improved generalization.

I am very confident in this model because the model was able to achieve 97% validation accuracy, showing strong generalization to unseen data.

Model deployment and inference: A brief description of how to deploy/serve your model and how to use the model for inference (this material should also be in the README with examples)

After training the CNN, the first step for deployment is to save the trained model, which I did using TensorFlow's built-in **model.save()** method. This stores the complete model, including the architecture, weights, and optimizer configuration, to a directory.

Once the model is saved, it can be deployed locally or in the cloud. I first tested to make sure the model worked by restarting the kernel and then testing the dataset on the saved model. In order to deploy it, I created a Flask server. The API would load the saved model and expose an endpoint that accepts image input and returns the model's prediction. When an image is sent to this endpoint, the server resizes and normalizes it before passing it through the model for inference. This setup allows me to integrate the model into applications or web services.

To perform inference with the trained model outside of a deployed service, I wrote a script that loads the model using **tf.keras.models.load_model()** and processes input images in the same way as during training, resizing them to the appropriate dimensions and scaling pixel values. The model then returns a confidence score, which can be interpreted directly, and its prediction.

All of this together, saving, serving, and using the model for inference that ensures that your machine learning pipeline is complete and ready for real-world use!

References:

Alternate LeNet-5 CNN: https://arxiv.org/pdf/1807.01688.pdf
Dropout Layer:
https://stackoverflow.com/questions/47892505/dropout-rate-guidance-for-hidden-layers-in-a-convolution-neural-network
ChatGPT:
For this project, I had a hard time deploying the interference server. Therefore, I used ChatGPT to help me with debugging the deployment. My main issues were that I already had something running on port 5000 and, therefore cannot deploy again. So I had to go in and terminate what was already running so that I could deploy again. Another issue that I had was that I was not able to pipe the images into my server. Therefore, ChatGPT helped me debug my **api.py** code so that it would be able to work with raw image files.