1. Provide your team background and organization description (if applicable).

I’m a data scientist based in Nairobi, Kenya, with a background in database management, data warehousing, and business intelligence. My journey in machine learning began in 2020 when I started exploring predictive modeling by participating in various data competitions. Over the past few years, I’ve competed in numerous challenges, leveraging my skills to tackle complex problems across diverse fields, including healthcare, satellite imagery, environmental conservation, and telecommunications. I’m passionate about continuous learning, contributing to data science competitions, and driving social impact through innovative solutions.

1. Explain why you participated in the DigiLut challenge.

I have participated in several digital pathology competitions before, and I wanted to see how my experience would translate to this new task. While past challenges primarily involved classifying whole slide images, this task required localizing a region of interest, which presented a different kind of challenge. I was curious to see if the techniques I had developed and refined in previous competitions would still be effective in this more focused task.

1. Describe how you built your winning model and elaborate on the technical and modeling choices you made.

**Data preprocessing**

Since the labeled regions represented only a small fraction of the dataset, I decided to create both positive and negative datasets to maximize the use of the available data.

For the positive dataset, I extracted regions around the labeled bounding boxes, adding random padding to introduce variability in the surrounding context. This approach helped the model generalize better by training it to recognize regions of interest even when the surrounding tissue varied. Through experimentation, I discovered that images extracted from page 3 provided a good balance between resolution and computational efficiency, resulting in a decent score without requiring excessive compute resources.

To create the negative dataset, I began by masking regions containing tissue at a low resolution (page 5). I then defined patches of various sizes, extracting the corresponding regions at page 3 to match the resolution of the positive dataset. Defining patches of various sizes in the negative dataset was an attempt to mirror the diversity in bounding box sizes found in the positive dataset. Next, I excluded any tiles that intersected with the labeled bounding boxes, ensuring that the negative dataset consisted entirely of non-target areas. This approach exposed the model to clear examples of irrelevant regions, enhancing its ability to distinguish between relevant and irrelevant areas during training.

**Model development**

I used a Faster R-CNN architecture with Extended-IoU (EIoU) loss to improve bounding box regression. I trained different models for each of the negative dataset sizes and employed 5-fold cross-validation. Finally, I combined predictions using Weighted Boxes Fusion (WBF) to aggregate the tile-level predictions from each model at the whole slide level. WBF was effective in merging overlapping predictions from different models, resulting in more accurate and reliable bounding box predictions.

**Ideas that didn’t work:**

* **Larger resolution images:** I experimented with using higher resolution images from pages 0, 1, and 2, but these proved too computationally intensive without providing significant improvements in accuracy.
* **Stain normalization:** I also explored stain normalization and augmentation techniques to reduce variability in slide staining, but these methods did not lead to better performance in this specific task.
* **Manual label correction:** Attempts to manually correct the labels in the dataset were unsuccessful. The effort to improve label accuracy didn’t translate into better performance on the test set, likely due to inconsistencies in the labeling process.
* **Extra datasets:** I attempted generating negative samples from the extra datasets. However, my experiments indicated a significant drift between the main dataset and the extra datasets, leading me to exclude the extra data from the final model. Although incorporating this extra data might have introduced more noise than value, it could still be valuable for pretraining.

1. What GPU/CPU/RAM resources you used to build your model

Tesla T4, P100 16GB from Google Colab and Kaggle.

1. Do you have any positive feedback or improvement opportunities for the Trustii.io platform?

The availability of basic compute resources on the platform was a significant advantage. It allowed me to preprocess data efficiently without needing to download it first, which greatly enhanced the accessibility of the competition.

For tasks that require a high level of specialization to assess label quality, ensuring clean and accurate labels is crucial. High-quality labels would enhance the reliability of the competition outcomes and provide a more robust foundation for model development.