Detection of Arsenic Contamination Using Satellite Imagery and Machine Learning

Agrawal, Ayush (School: Canyon Crest Academy)

Arsenic, a WHO-classified carcinogen and neurotoxin, has gained widespread notoriety, potentially affecting 250+ million globally. Current arsenic detection methods require frequent sampling, expensive equipment, and trained personnel, being difficult to implement in developing regions and especially during COVID-19. This study attempts to establish a relationship between soil's hyperspectral satellite data and arsenic content. Four regression-based machine learning models such as Random Forest (RF) are tested to determine this correlation from NASA's Hyperion satellite, with linear regression as a control. Raw data is converted to reflectance, problematic atmospheric windows are removed, and 4 noise reduction algorithms are tested. Dimensionality reduction, data augmentation (DA) techniques, hyperparameter optimization, and overfitting are addressed. Validation results indicate success for the SD+DA+RF model (R2 = 0.872 and RMSE= 0.091), establishing a strong proof-of-concept despite limited data. Three multi-class classification machine learning models are then applied to characterize regions in the western United States with 3 tiers of contamination risk, achieving up to 76% accuracy with 4 different land covers (bare, shrubland, grassland, agricultural). Future work involves establishing more holistic contamination risk criteria. Overall, these results, being the first to use hyperspectral satellite-based data, suggest that such a methodology is scientifically practical and implementable, and can provide a low-cost, rapid, less labor-intensive, and scalable alternative to arsenic contamination detection. The author is currently engaged with professional initiatives to further the tool's accessibility.

Awards Won:

Second Award of $2,000