Preventing Deforestation: Modeling and Prediction of Vulnerabilities in Forest Conservation

Saptarashmi Bandyopadhyay [*1], **Deepthi Raghunandan** [*1], **Dhruva Sahrawat** [1], **John Dickerson** [1]

[1] Department of Computer Science
University of Maryland, College Park
Maryland, MD 20742
saptab1@umd.edu, draghun1@umd.edu, dhruva7@umd.edu, johnd@umd.edu

Abstract

We predict attacks on tree cover, a green security asset, in subnational regions of Indonesia using a boosted Decision Tree Classifier, the BoostIT algorithm. Our models are based on a thorough literature survey which found that deforestation occurs in hotspots, and proximity to other anthropomorphic activity is the strongest predictor of deforestation in other subnational regions. Coarse-grained prediction of targets vulnerable to attacks is a significant challenge in Green Security Games for strategizing by defenders. We find that a boosted Decision Tree Classifier takes minimal resources to build, is accurate in its predictions, and is scalable for the sake of expanding on the assumptions made regarding the drivers of deforestation. We show that such an algorithm can empower communities to manage forest resources effectively.

1 Introduction

Accurately detecting and predicting *hotspots* of vulnerable forest assets is a necessary but challenging task in the domain of *green security* problems, protection of natural assets subject to strategic adversaries (see, e.g., Fang and Nguyen 2016). Solutions in this space are applied towards the management of limited forest resources and, ultimately, the health of our planet. Literature suggests that government bodies, who are tasked with managing resource-rich forests, rarely have comprehensive data regarding deforestation in their region (Austin et al. 2019). This is due, in part, to the difficulty of accurate data collection for those protecting green assets, and the quantity of data—both in terms of its temporal and spatial depth—can be overwhelming for the many stakeholders involved.

We believe such predictive models can empower stakeholders by pointing them to the nexus of human-driven deforestation. We focus on applying predictive models specifically to the Global Forest Watch [1] data from Indonesia. The reason why we considered Indonesia was because deforestation is occurring on a massive scale there—ripe for modeling. As we see in Figure 1, deforestation in Indonesia has accelerated in severity in the last 20 years. This data is readily available. Previous works have been able to detect upwards of 5 different drivers of forest loss using satellite imagery alone (Austin et al. 2019).

We view deforestation as *adversarial behavior* against forest assets by attackers in a green security game framework (Fang, Stone, and Tambe 2015). This lens enabled us to model deforestation as it occurs—much like criminal activity—at *hotspots* which shift over time. We view "attackers" as individuals who extensively consume forest assets and "defenders" as those individuals who work to preserve those same assets.

We see that models need to predict the extent of demand for available resources and the growth rate of renewable assets. These types of predictors could encourage stakeholders to implement forest-farming practices; in that they reap only the assets which meet current needs and reduce waste. In generating fast, efficient, and detailed prediction models for the short term and long term, we may support the development of better mechanisms to empower all people to consider the actual cost of exploiting their resources.

In this work, we focus on specifically predicting the vulnerable areas pertaining to green assets. These predicted vulnerabilities can be used as input data to enhance the decision-making in green security problems. To provide insight into curbing forest loss, we propose a predictive model. We use tree-loss-cover data collected over a 20-year period to define attacks on forest assets. We categorize an area as being attacked when there is a net-positive percent of tree-cover-loss. Based on a map of attacks within different sub-national regions in Indonesia, we built a spatially-aware boosted decision tree model, with iterable learning (BoostIT), to predict hotspots vulnerable to deforestation by attackers. We boosted our classification using data regarding geographic parameters (elevation in particular) to determine the accessibility of particular regions. We evaluated the performance of the pure decision tree model (with and without iterable learning), and the performance of the decision tree model with geographic parameters (with and without iterable learning). We found that all four of our classifiers performed well and gave us some idea of which areas are most vulnerable to deforestation.

[*] first authors

[1] https://www.globalforestwatch.org/

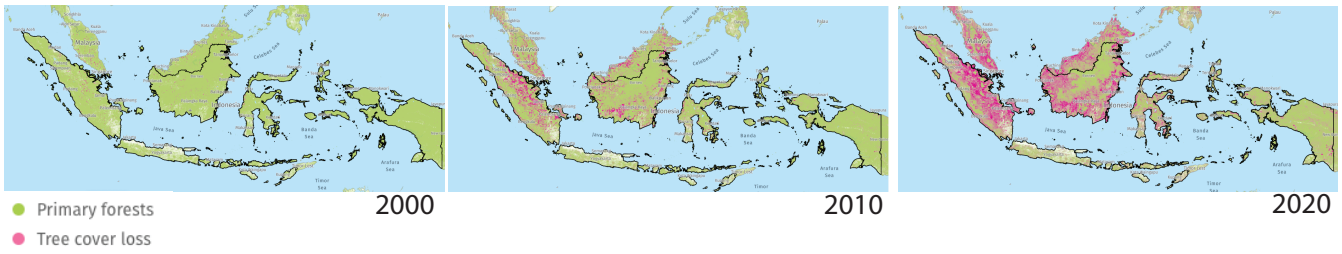

Figure 1: We are focusing on tree cover loss, specifically in Indonesia, due to the severely vulnerable status of forest assets in Indonesia over the past 20 years. It can be seen in the image corresponding to the year 2000 that there was a minor tree cover loss just 20 years ago. The image corresponding to 2010 shows a moderate loss in the Western islands of Sumatra with not much change in the remaining islands of Indonesia. However, the tree cover loss started to increase from 2010- 2020 rapidly.

2 Related Work

We observed that environmental science and mechanism design literature addressed our problem-space most directly. In the following sections, we combine the findings from both fields to make a case for the type of solution we prototype.

Three leading studies (Austin et al. 2019; Wade et al. 2020; Irvin et al. 2020) conducted a large-scale analysis of geographic images to identify trends in deforestation. All three projects conduct retrospective land-cover analysis to correlate in-situ data with satellite imagery to identify drivers.

Wade et al. (2020) performed a study in 2018 which specifically identified the drivers in the protected areas of Indonesia. They found a high correlation between the tree cover loss in protected areas and agricultural expansion. In Indonesia, fires started by farmers near protected lands often unintentionally destroyed peat forests. The growth of scrubland, detected on satellite images, followed by forest loss, helped confirm this theory—more than 38% - 57% of all protected areas which were deforested were left to become scrubland or grassland.

Austin et al. (2019) used similar techniques to understand the drivers of deforestation in all of Indonesia. They found that between 2001 and 2016, oil palm plantations were the single largest contributor, accounting for 23% of all deforestation across the country. Though this is still one of the primary drivers of forest loss, deforestation from the oil palm industry hit a peak between 2008 and 2009 and has been steadily declining. Austin et al. (2019) also identified other drivers, ordered from highest to lowest impact: timber industry, unintentional conversion of forests to grassland, small-scale agriculture, small-scale clearing, mining, and fish ponds.

The latest study, conducted by Irvin et al. (2020), introduced a unique deep-learning technique to automatically identify when deforestation is taking place and its causes. As with previous studies, ForestNet learned to identify drivers by using expertly labeled satellite images. The main thrust of this work is not in identifying unique drivers but in demonstrating that an automatic approach can be employed to identify when these drivers are affecting changes in Indonesian forests. However, since the images being labeled are static and a proper analysis of human drivers requires observation of trends, it is unlikely that this approach has real-world applications. Nevertheless, it is a start for those who want to build models which provide more granular predictions like our own!

Two of the papers we surveyed sought to predict the land areas which were susceptible to deforestation (Kayet et al. 2021; Gaveau et al. 2021). Kayet et al. (2021) developed a spatial-temporal analysis technique to identify the primary drivers of deforestation in the Saranda forest of India. They used data within a *GIS framework* to make predictions about the land segments which were the most susceptible to deforestation into the year 2050. Using a combination of analysis techniques, including frequency ratio, logistic regressions, and an analytic hierarchy process, they showed that the susceptible areas were the most likely to be in proximity with settlements or current anthropomorphic activities. This finding is consistent with the literature, which models poaching or illegal logging behavior using *crime hotspots* (Xu et al. 2020; Mc Carthy et al. 2016; Johnson, Fang, and Tambe 2012).

Gaveau et al. (2021) observe and predict trends specifically in Indonesian New Guinea. They study the relationship between the proximity of forests to new or existing roads and accessibility based on land properties (slope, elevation, and cost distance to public roads). Weights on existing evidence and logistic regression are used to generate a map that demarcates the areas susceptible to deforestation. Both these papers ultimately model *hotspots* of forest activity, finding that proximity to previous anthropomorphic activity drives current trends.

Green Security Games can be conceptualized as a game between attackers and defenders on particular targets holding high significance in green security with coverage vectors indicating if the targets have been covered (Nair 2018). There can be complex versions of this game played over several rounds in order to achieve equilibrium.

We found a few different projects which were built on

this premise and applied towards predicting illegal logging (Mc Carthy et al. 2016; Johnson, Fang, and Tambe 2012), and poaching (Xu et al. 2020). These works primarily focus on finding the optimum patrol/defender strategy to prevent illegal extraction of forest resources either by using multi-arm bandits (Xu et al. 2020), neural networks, or by modeling it as a Stackelberg game (Mc Carthy et al. 2016) (Johnson, Fang, and Tambe 2012). Instead of focusing on optimum defender strategy, we focus on a smaller part of this problem by finding and predicting the most vulnerable forest assets based on previous forest cover loss data and other domain features.

However, for our purpose, the Kar (2017) thesis introduces a spatially-aware BoostIT (boosted decision tree with an iterative learning algorithm), which we plan to use in our deforestation problem. Their thesis uses crime hotspots with soft boundaries in the decision space for less fine-grained segmentation and, therefore, more capable of representing hotspots to protect wildlife. We plan to compute distance among hotspots with reduced forest cover like their distance function does. They use a parameter vector λ to get the importance of all features, a parameter ω to measure the impact of domain features on observation probability and a β parameter to estimate detector efficiency. These could be interesting for our model. We further motivate our work with a detailed literature review in Appendix A.

3 Dataset

We acquired our dataset through the Global Forest Watch Initiative. The data between 2000–2013 was collected and verified, in blocks, by Hansen et al. (2013). Since 2013, the global multi-spectral observations were collected by Thematic Mapper Plus (ETM+) onboard Landsat 7, and the Operational Land Imager (OLI) onboard Landsat 8 at a 30m spatial resolution.[2] These images were used to produce the tree cover dataset—a product of the University of Maryland's GLAD lab and Google. They enriched image recognition with *in-situ* data—collected, on the ground, by environmental scientists. Additionally, we have computed terrain features like slope using the Google Earth Engine Python API.[3] Elevation data from the Shuttle Radar Topography Mission (SRTM) (Farr et al. 2007) was processed and made accessible by NASA JPL at a resolution of 30m.

Our primary dataset contained five major aspects of deforestation and climate change, namely: (1) forest loss, (2) biomass loss, (3) CO_2 emissions, (4) different densities of canopy cover, and (5) sub-national regions in Indonesia (henceforth 'subnational1' and 'subnational2' are referred to as a state and a district respectively). We computed the forest loss data from available Forest Change Data over ten years and the percentage of land deforested compared to the total land area for districts in a state. The percentage of deforested land determined whether forests are being *attacked* or not. Please refer to Appendix B for a detailed description of our parameters, our labeling methods, and the generation of the train-test data-splits as inputs for model construction.

[2]https://glad.umd.edu/dataset/glad-forest-alerts

[3]https://github.com/google/earthengine-api

4 Methods

We implemented our version of the spatially aware BoostIT algorithm, introduced by Kar (2017) to identify areas that are vulnerable to forest loss. We view loggers as attackers of forests. The goal for this green security domain is to protect the green assets from attackers. Unlike wildlife, which can be targeted in protected areas like national parks, trees can be anywhere. Thus, the notion of defenders in this green security game can and should be broad. We will recommend probable defenders who can effectively act on information on vulnerable areas prone to deforestation, subject to incentives.

Initially, we trained a Decision Tree Classifier to predict areas vulnerable to tree-cover loss using features from the original dataset. We improved it via the BoostIT decision tree algorithm as outlined in Algorithm 1. We predicted hotspot labels on training data and test data.

Next, to calculate hotspot proximity, we first devised Algorithm 2 to compute our notion of proximity in the dataset. The simple idea of proximity is that all districts in a state are proximate, but districts in different states are not. This algorithm considers predicted hotspot class labels from training data (Θ^h) or test data (Ψ^h) as inputs. Then it finds their state ids denoting whether they are close to one another. Then for each district, the number of districts close to it is computed. If this number is above a certain threshold α, a new spatial feature (h) is added to the dataset updating the training data (Θ^h) or test data (Ψ^h) accordingly—thereby learning the boosted decision tree. α is set to 5 in our experiments. This was repeated for a fixed number of iterations (10 times in our experiment). We used the original decision tree and the ten boosted decision trees for accuracy calculation and further analysis.

Please refer to Appendix C for details regarding hotspot proximity calculation and to see the visualization of our boosted decision tree model (Figure 3).

5 Results

The results on the comparison between different models are presented in Table 1.[4] We can see that all four models have decent performance (accuracy $> 62\%$) and can give some idea of which areas are most vulnerable to deforestation. The mismatch in the precision and recall between the two labels can be explained by the fact that even if two locations have the same features, they could be labeled differently. Boosting the decision tree with proximity lead to significant improvement in prediction results—both for the base model and the base model with the addition of terrain features (see Figure 3). This approach aligns with the intuition that identifying areas vulnerable to deforestation is more akin to detecting hotspots than segmentation. The addition of terrain features also leads to a significant jump in performance. Biomass features seem to be the most important for our base model. However, the slope and biomass standard

[4]We have defined each model and included the confusion matrix of the boosted base model with terrain features in Figure 2, in Appendix D.

Model	Accuracy	Vulnerable		Not Vulnerable	
		Precision	Recall	Precision	Recall
Base model	62%	72%	76%	27%	23%
Base model with BoostIT	67%	79%	77%	35%	32%
Base model with terrain features	69%	77%	79%	51%	48%
Base model with terrain features and BoostIT	73%	80%	83%	59%	55%

Table 1: Performance of different versions of our model on the test data.

Algorithm 1: BoostIT algorithm to detect targets vulnerable to deforestation

Input: $train\_data$, $test\_data$, $proximity\_vector$, $iterations$
Parameter: $alpha = 5$
Output: $trees$, $gtrlist$, $gtelist$

$\Theta_0 \leftarrow train\_data$
$\Psi_0 \leftarrow test\_data$
$D_0 \leftarrow learn\_decision\_tree(\Theta_0)$
$gtr_0 \leftarrow predict\_labels(D_0, \Theta_0)$
▷ Predict labels on training data
$gte_0 \leftarrow predict\_labels(D_0, \Psi_0)$
▷ Predict labels on test data
$i \leftarrow 0$
$trees \leftarrow$ initialized as list with 1st element as D_0
$gtrlist \leftarrow$ initialized as list with 1st element as gtr_0
$gtelist \leftarrow$ initialized as list with 1st element as gte_0
$\Theta_0^h \leftarrow \Theta_0$
$\Psi_0^h \leftarrow \Psi_0$
while $i < iterations$ **do**
  $htrain\_name \leftarrow$ "$h\_train\_iteration$" $+ str(i)$
  $\Theta_i^h \leftarrow calc\_hotspot\_prox(gtr_i, \Theta_i^h, htrain\_name)$
  ▷ Spatial feature is added to training features
  $htest\_name \leftarrow$ "$h\_test\_iteration$" $+ str(i)$
  $\Psi_i^h \leftarrow calc\_hotspot\_prox(gte_i, \Psi_i^h, htest\_name)$
  ▷ Spatial feature is added to test features
  $D_i \leftarrow learn\_decision\_tree(\Theta_i)$
  $gtr_i \leftarrow predict\_labels(D_i, \Theta_i)$
  ▷ Predict labels on training data
  $gte_i \leftarrow predict\_labels(D_i, \Psi_i)$
  ▷ Predict labels on test data
  $trees$.append(D_i)
  $gtrlist$.append(gtr_i)
  $gtelist$.append(gte_i)
  $i \leftarrow i + 1$
return $trees, gtrlist, gtelist$

deviations for the base model, with terrain features, were equally essential features.

6 Conclusion

We have noticed an improvement in accuracy, precision, and recall metrics from the base decision tree model to the BoostIT algorithm adapted to predict vulnerable regions susceptible to deforestation. The precision of predicting regions vulnerable to deforestation with the basic decision tree classifier is 72%, improving to 79% once the BoostIT algorithm is implemented. The results improve further with terrain features like slope, reaching even 80% precision in detecting vulnerable hotspots, which indicate its importance where elevation can be used to transfer logged trees by pushing them down from a hill. Our BoostIT decision tree algorithm can be extended to a spatial version of XGBoost (Extreme Gradient Boosting) (Chen et al. 2015) with the aim of improving prediction accuracy.

Our results indicate that we can sustain and improve good model performance in detecting regions vulnerable to forest loss with the addition of features. We believe that this implies that generating reliable data using automatic techniques as input for green security games is very achievable. Our current approach aimed to show that even simple models can efficiently and accurately predict vulnerable regions. Our algorithm can be easily applied to any region around the world so long as forest-cover loss data is available.

Game-theoretic models, including the green security game paradigm (Fang, Stone, and Tambe 2015), are only as good as their input data. We address that first stage of the pipeline. We anticipate that generating uncertainty metrics of our predictions will help to enhance decision-making within the green security game (Chipman, George, and McCulloch 2010).

Future work could extend our approach by considering additional fine-grained features like the proximity of vulnerable hotspots to rivers, roads, human settlements, and farmland obtained from Google Geo-location services using satellite images. One potential drawback is that as soon as we consider these features, the variance of the data will likely increase drastically. Forest cover will depend upon features which may be hard to obtain like the ownership of land and the distribution of trees in those areas.

Also, the features can change over time. For example, new roads can be built, and topography can change due to natural disasters. We consider forest loss at a granular district level because that helps in the precise identification of hotspots and consequently better enables the implementation of concrete strategies towards protecting vulnerable green assets.

Our project on the necessity of detecting vulnerable hotspots susceptible to deforestation can be the stepping stone for green security games and policy work to create a foundation for the protection and management of forest resources. Unlike defenders protecting wildlife assets in specific conservation areas, defending against illegal logging and managing forest assets is challenging as deforestation is a widespread, pervasive and global problem.

References

Abram, N. K.; Meijaard, E.; Wilson, K. A.; Davis, J. T.; Wells, J. A.; Ancrenaz, M.; Budiharta, S.; Durrant, A.; Fakhruzzi, A.; Runting, R. K.; Gaveau, D.; and Mengersen, K. 2017. Oil palm–community conflict mapping in Indonesia: A case for better community liaison in planning for development initiatives. *Applied Geography*, 78: 33–44.

Austin, K. G.; Schwantes, A.; Gu, Y.; and Kasibhatla, P. S. 2019. What causes deforestation in Indonesia? *Environmental Research Letters*, 14(2): 024007.

Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4): 1–4.

Chipman, H. A.; George, E. I.; and McCulloch, R. E. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

Fang, F.; and Nguyen, T. H. 2016. Green security games: Apply game theory to addressing green security challenges. *ACM SIGecom Exchanges*, 15(1): 78–83.

Fang, F.; Stone, P.; and Tambe, M. 2015. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Farr, T. G.; Rosen, P. A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. 2007. The shuttle radar topography mission. *Reviews of geophysics*, 45(2).

Friedman, R. S. 2020. Managing forests for community, conservation, and social equity: a case study of social forestry in Indonesia.

Gaveau, D.; Santos, L.; Locatelli, B.; Salim, M.; Msi, H.; Meijaard, E.; Heatubun, C.; and Sheil, D. 2021. Forest loss in Indonesian New Guinea: trends, drivers, and outlook.

Hansen, M. C.; Potapov, P. V.; Moore, R.; Hancher, M.; Turubanova, S. A.; Tyukavina, A.; Thau, D.; Stehman, S. V.; Goetz, S. J.; Loveland, T. R.; Kommareddy, A.; Egorov, A.; Chini, L.; Justice, C. O.; and Townshend, J. R. G. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160): 850–853.

Irvin, J.; Sheng, H.; Ramachandran, N.; Johnson-Yu, S.; Zhou, S.; Story, K.; Rustowicz, R.; Elsworth, C.; Austin, K.; and Ng, A. Y. 2020. ForestNet: Classifying Drivers of Deforestation in Indonesia using Deep Learning on Satellite Imagery. *arXiv preprint arXiv:2011.05479*.

Johnson, M.; Fang, F.; and Tambe, M. 2012. Patrol strategies to maximize pristine forest area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

Kar, D. 2017. *When AI helps Wildlife Conservation: Learning Adversary Behaviors in Green Security Games*. Ph.D. thesis, University of Southern California.

Kayet, N.; Pathak, K.; Kumar, S.; Singh, C.; Chowdary, V.; Chakrabarty, A.; Sinha, N.; Shaik, I.; and Ghosh, A. 2021. Deforestation susceptibility assessment and prediction in hilltop mining-affected forest region. *Journal of Environmental Management*, 289: 112504.

Mc Carthy, S. M.; Tambe, M.; Kiekintveld, C.; Gore, M.; and Killion, A. 2016. Preventing illegal logging: Simultaneous optimization of resource teams and tactics for security. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Monzon, J. P.; Slingerland, M. A.; Rahutomo, S.; Agus, F.; Oberthür, T.; Andrade, J. F.; Couëdel, A.; Edreira, J. I. R.; Hekman, W.; van den Beuken, R.; et al. 2021. Fostering a climate-smart intensification for oil palm. *Nature Sustainability*, 1–7.

Nair, S. 2018. When Security Games Go Green. https://www.cs.umd.edu/class/spring2018/cmsc828m/lecs/cmsc828m_s2018_lec9_suraj_nair.pdf. [Online; accessed 9-April-2021].

Santika, T.; Meijaard, E.; Budiharta, S.; Law, E. A.; Kusworo, A.; Hutabarat, J. A.; Indrawan, T. P.; Struebig, M.; Raharjo, S.; Huda, I.; Sulhani; Ekaputri, A. D.; Trison, S.; Stigner, M.; and Wilson, K. A. 2017. Community forest management in Indonesia: Avoided deforestation in the context of anthropogenic and climate complexities. *Global Environmental Change*, 46: 60–71.

Wade, C. M.; Austin, K. G.; Cajka, J.; Lapidus, D.; Everett, K. H.; Galperin, D.; Maynard, R.; and Sobel, A. 2020. What is Threatening Forests in Protected Areas? A Global Assessment of Deforestation in Protected Areas, 2001–2018. *Forests*, 11(5): 539.

Xu, L.; Bondi, E.; Fang, F.; Perrault, A.; Wang, K.; and Tambe, M. 2020. Dual-Mandate Patrols: Multi-Armed Bandits for Green Security. *arXiv preprint arXiv:2009.06560*.

Zarin, D. J.; Harris, N. L.; Baccini, A.; Aksenov, D.; Hansen, M. C.; Azevedo-Ramos, C.; Azevedo, T.; Margono, B. A.; Alencar, A. C.; Gabris, C.; et al. 2016. Can carbon emissions from tropical deforestation drop by 50% in 5 years? *Global change biology*, 22(4): 1336–1347.

Appendix A Literature Survey

Preventing Deforestation

Other Potential Mechanisms We think that future work in this area would empower local stakeholders to manage forest resources and make it the responsibility of the international community– as they also reap the benefits of these resources. We found quite a few papers in environmental science and economics that proposed solutions towards this end. One solution would be to cap and enforce carbon emissions of international countries tasked with managing resource-rich tropical forests (Zarin et al. 2016). This approach would limit the amount of deforestation those countries could leverage towards economic gain. However, this can also cause countries to refuse to play in that market—particularly if they think they can leverage their resources for monetary benefits elsewhere.

It seems that the real issue is that communities tasked with caring for rich forest resources are often at an economic disadvantage in international markets. They find that it is advantageous to sell their resources to powerful multinationals to elevate the status of their country and its people. Thus, mechanisms applied towards reflecting the actual cost of using these resources, be it at the local or global scale, must elevate the voices and power of these local communities (Friedman 2020). This means providing detailed information to local communities regarding the resources available on their land and the best way to harvest those resources without creating permanent damage to the forest itself. We found lots of literature that confirmed these ideas and proposed mechanisms to that end. For example, a few proposed enabling a community forestry scheme to allow local communities to directly manage forest resources (Santika et al. 2017; Abram et al. 2017). Monzon et al. proposed, in a nature article, a scheme towards bettering the management of existing plantations in order to reduce deforestation (Monzon et al. 2021).

To this end, algorithms which reflect these values must be: (1) transparent, (2) readily accessible to local communities, and (3) include information on all the facets which influence deforestation. Details could include: availability of specific resources in a region, the accessibility of those resources across all regions, how these resources are actually valued in the international market, and the risks involved (to bio-diversity or the health of the forests) in harvested those resources.

The ease with which we can train a decision tree classifier and the extensibility of this approach warrant an exploration into using decision tree classifiers towards empowering community-driven forest-management initiatives.

Appendix B Dataset

Parameters and Features

We considered some basic attributes as parameters for our model. Some of these attributes were unique to Indonesia, and some are general to measuring hotspots of deforestation in any Green Security domain. Our primary dataset contained five major aspects of deforestation and climate change, namely:

- tree loss cover measured in hectares over 20 years
- biomass loss measured in metric tonnes over 20 years
- CO_2 emissions measured in metric tonnes over 20 years
- different densities of canopy cover measured by the percentage of the tree cover over the land
- and sub-national regions in Indonesia which are analogous to states and districts.

Tree cover loss data has been critical to label sub-national regions vulnerable to deforestation. Tree cover is defined as all vegetation greater than 5 meters in height. It can take the form of natural forests or plantations across a range of canopy densities. Biomass and CO_2 emissions information were calculated by Hansen et al. They correlated in-situ data with image detection techniques to get these attributes. Thresholds of canopy cover ($\geq$ 10%, 15%, 20%, 30%, 50%, and 75%) measure the density of trees in a sub-national region. Given a particular area, represented as a circle, the percent canopy cover statistic relates to the percentage of the canopy's ground area. We looked at data produced between 2000 and 2020 across all canopy densities and all sub-national regions.

Based on existing literature and the data available to us, we considered and used the set of features to build our spatially aware boosted decision tree model, with iterable learning (BoostIT), to predict hotspots vulnerable to deforestation by adversaries. We boosted our classification using data regarding geographic parameters (elevation in particular) to determine the accessibility of particular regions. We acknowledge that this is a rough estimate of accessibility. In addition, we considered metric tonnes of CO_2 emissions and metric tonnes of loss in biomass to reinforce that an attack had taken place. Our prediction model specifically aims to predict vulnerable regions susceptible to palm tree-related land clearing ((Austin et al. 2019)).

Labeling Tree Cover Loss as Attacks We computed the forest loss data from available Forest Change Data over ten years and computed the percentage of land deforested compared to the total land area in 'subnational2' (henceforth referred to as district) in a 'subnational1' region (henceforth referred to as state). The percentage of deforested land determines whether forests are being *attacked* or not. This will be our dependent variable.

If the percentage of deforested land is negative, it indicates that afforestation has happened in that district. Therefore, it is assigned a class label of 0, indicating that the forests in that district for the particular canopy threshold are not vulnerable to deforestation. If the percentage of deforested land is positive, it indicates that deforestation has happened in that district; therefore, it is assigned a class label of 1 indicating that the forests in that district for the particular canopy threshold are vulnerable to deforestation. For reliability, the annotation of data for targets vulnerable to deforestation must be professionally done by climate change forest experts. Here, they have been automatically labeled for analysis on a low scale with limited resources.

Creation of Training and Test Data In the original dataset, there were 3514 data elements, comprising of 502

districts with seven canopy thresholds. Overall there are 33 distinct states in the datasets. We refer to each data element based on ('subnational1', 'subnational2') keys equivalent to (state, district) as two states can have the same district name.

We split the available Indonesian forest loss data from Global Forest Change in an 80/20 split to training datasets and test datasets by ensuring that the distribution of the class labels (24% negative labels and 76% positive labels) in the original dataset is maintained in both the train and test split. Suppose we are taking data for a (state, district) with a particular canopy threshold in the training file. In that case, data for that same (state, district) for the remaining six canopy thresholds are also placed in the training dataset. The same step is repeated for test data. This is done to have a proper separation between test data and training data.

Appendix C Methods

Hotspot Proximity Calculation To calculate hotspot proximity, we have first devised an algorithm to compute our notion of proximity in the dataset as outlined in algorithm 2. The simple idea of proximity is that all districts in every state are proximate to one another, but districts in different states are not. This oversimplification may be pertinent in Indonesia, with thousands of islands formed states based on proximity. Further research can be done to calculate the exact distance between districts to make bordering districts in adjoining states proximate to one another.

The proximity vector is computed, which is an array of 3514 elements as there are 3514 data points. Each element is the state id for the particular district. The notion of proximity as outlined in algorithm 2 is essential to calculate hotspot proximity in algorithm 3 which is used in the BoostIT algorithm.

This algorithm considers predicted hotspot class labels from training data (Θ^h) or test data (Ψ^h) as inputs. Then it finds their state ids denoting whether they are close to one another. Then for one district, the number of districts close to it is computed. If this number is above a certain threshold α, a new spatial feature (h) with value one is added to the dataset updating the training data (Θ^h) or test data (Ψ^h) accordingly. α is set to 5 in our experiments.

Appendix D Results

We evaluate the following four versions of our model on the test split of our dataset in order to understand how do our assumptions affect the performance of the model :

- **Base model**: This model is composed of a decision tree used as a classifier.
- **Base model with BoostIT**: This model is composed of a decision tree along with the BoostIT algorithm applied to it as shown in Algorithm 3.
- **Base model with terrain features**: This model is composed of a decision tree with the inclusion of terrain features (e.g. slope).
- **Base model with terrain features and BoostIT**: This model is composed of a decision tree with the inclusion of terrain features (e.g. slope) and also with the BoostIT algorithm applied to it as shown in Algorithm 3.

Algorithm 2: Proximity Calculation

Input: $orig\_data$
Output: $proximity\_vector$

$l \leftarrow length(orig\_data)$
▷ orig_data has features for every canopy threshold in every district of every Indonesian state
$proximity\_vector \leftarrow$ initialized to an array with 0s having length l
$unique\_states \leftarrow orig\_data['subnational1'].unique()$
▷ Gets a list of Indonesian states
$state\_id \leftarrow 0$
for $each\_state$ in $unique\_states$ **do**
    $index\_all\_districts\_in\_a\_state \leftarrow$ list of all indices of $each\_state$ in $orig\_data$
    for $index$ in $index\_all\_districts\_in\_a\_state$ **do**
        $proximity\_vector[index] = state\_id$
    $state\_id \leftarrow state\_id + 1$
return $proximity\_vector$

Algorithm 3: Calculate Hotspot Proximity

Input: $predictions$, $feature_name$, $data$
Parameter: $alpha = 5$
Output: $data$

$hotspot\_indices \leftarrow$ gets a list of indices in $data$ for hotspots (class label = 1) in $pred$
$hotspots \leftarrow$ state_ids from $proximity\_vector$ at $hotspot\_indices$
$hotspot\_state\_counts \leftarrow$ a 2D array with unique states and their counts in $hotspots$
$data[feat\_name] \leftarrow 0$
▷ A new column initialized to 0 is added to the $data$
for $state\_count$ in $hotspot\_state\_counts$ **do**
    $state = state\_count[0]$
    $count = state\_count[1]$
    if $count \leq alpha$ **then**
        $dist\_indices\_cutoff \leftarrow$ indices from $prox\_vec$ for state_ids with value $state$
        $data[dist\_indices\_cutoff, feat\_name] = 1$
▷ Sets the new feature in $data$ as 1 for elements in $data$ at $dist\_indices\_cutoff$
return $data$

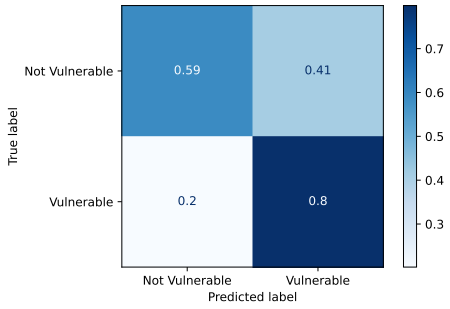

Figure 2: Confusion matrix for the base model with terrain features and BoostIT

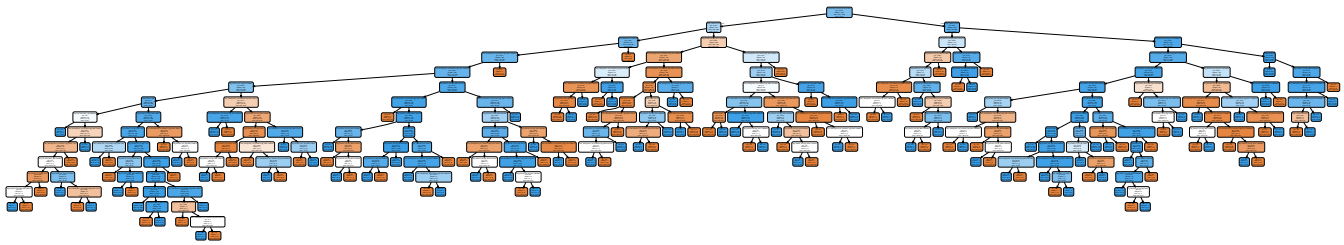

Figure 3: Visualization of the decision tree learned for the base model with terrain features and BoostIT (Please zoom in to see the individual nodes of the decision tree more clearly)