DETECTION OF DEFORESTATION USING SATELLITE IMAGES IN SRI LANKA

A data science project report for the course
"Machine Learning - (DSC 4173)"
presented by

B.P.N. BALASOORIYA, (S/17/317)

K.M.S.S.B. KULASEKARA, (S/17/403)

K.A.L.S. KULASOORIYA, (S/17/404)

W.G.K.D. WEERASINGHE, (S/17/509)

DEPARTMENT OF STATISTICS & COMPUTER SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF PERADENIYA
SRI LANKA
2023

ABSTRACT

Deforestation is a major concern in many parts of the world, including Sri Lanka. To effectively address this issue and develop appropriate conservation strategies, it is crucial to accurately detect and monitor deforestation. One approach to detect deforestation is through the use of satellite imagery, specifically Landsat satellite images and the analysis of their Normalized Difference Vegetation Index values. Therefore, this study is to detect the changes in forest cover in Horupathana National Park during the period from 2020 to 2023 using ARIMA and Random Forest models. According to the data validation, the Mean Absolute Error 0.297, Mean Squared Error 0.154 and the R-squared Score: 0.9982. As a result, the forest cover decreased to 18.864% compared to 2020 levels. Legal measures should be pursued by the relevant authorities to address the escalating deforestation issue, and a comprehensive approach is crucial to safeguard the diminishing forest resources within the research region.

KEYWORDS: Deforestation, Satellite image, ARIMA, Random Forest

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our project supervisor Dr.Mahasen Dehideniya for his invaluable guidance and support throughout the duration of this project. His patience and expertise have been instrumental in helping us navigate the challenges and complexities of this project. Finally, we would like to express our appreciation to our family and friends for their unwavering support and encouragement during this project. Their love and understanding have been a constant source of motivation and inspiration.

TABLE OF CONTENT

SECTION 1: Introduction

1.1. Importance

Deforestation is considered as a global problem and there have been many steps taken worldwide to control it. The devastation of forests and natural resources compromises the ecological balance and seriously affects the economy and quality of life across the planet by causing the loss of biodiversity, climate change, habitat loss, increased global warming, and other problems. According to the Global Forest Watch, from 2001 to 2022, there was a total of 459 million hectares of tree cover loss globally, which equates to a 12% decrease in tree cover since 2000. The World Resources Institute states that deforestation increased 12% globally in 2020. One way to keep an eye on deforestation is by using remote sensing. This is about getting pictures of the Earth's surface from far away using special cameras called remote sensors. These pictures come from devices on satellites in space.

According to those global rates, deforestation is a growing concern worldwide, and Sri Lanka is no exception to this environmental challenge. Sri Lanka is a tropical island nation located in the Indian Ocean, Sri Lanka is home to lush rainforests and diverse ecosystems that provide critical habitats for numerous plant and animal species. Sri Lanka has lost 210,000 hectares (520,000 acres) of tree cover since 2001. This means that Sri Lanka has lost 5.3% of its tree cover since 2000 by leading to a 1.14% deforestation rate per year. This has become a significant threat to Sri Lanka’s natural heritage. The main drivers of deforestation in Sri Lanka are agriculture, infrastructure development, and logging. Agriculture is the leading driver, accounting for about 60% of deforestation. Infrastructure development is responsible for about 20% of deforestation, and logging accounts for about 10%.

Deforestation in Horowpathana National Park, Sri Lanka, is a pressing issue that requires effective monitoring and detection methods. One approach to detecting deforestation in this region is through the use of Landsat satellite images and the analysis of their Normalized Difference Vegetation Index values. The Normalized Difference Vegetation Index is a valuable remote sensing tool that provides information about vegetation activity and health.

.By analyzing the NDVI values of Landsat satellite images, we can assess changes in vegetation cover over time and identify areas that have experienced deforestation (Chiteculo et al., 2018, #).Our project is all about using these satellite pictures and a special number called NDVI to figure out if deforestation is happening and help us address the pressing issue of deforestation

1.2. Problem Statement

Find the deforestation that happened in Horowpathana National Park during the period of 2019– 2023 and how can we obtain a model to predict deforestation using a machine learning algorithm?

1.3. Motivation

Over the last couple of years, the issue of deforestation in Sri Lanka has been widely discussed in the media. This has led to increased awareness among the public and in political circles. Additionally, there have been petitions raised against the cutting down of forests. Hence, finding a solid solution to forest management is crucial. This helps keep our environment safe and makes sure all the different plants and animals in our country stay healthy.

SECTION 2: Literature Review

From ancient times trees have been cut down for different purposes. Cutting down trees in large amounts is known as deforestation (Liyanawatte & Dias, n.d.). Deforestation can happen based on many factors like colonization, agriculture, logging, Infrastructure, war etc. (Serasinghe, n.d.) Previous studies have demonstrated the effectiveness of using Landsat satellite images and NDVI analysis for detecting deforestation and forest degradation in various regions.

For example, studies conducted in Brazil have successfully used Landsat images and NDVI analysis to identify deforestation and forest degradation over a span of 10 years. The deforestation contribution to carbon emission is already known but determining the contribution of forest degradation remains a challenge (Shimabukuro et al., 2019, #).

These studies have also explored the use of Landsat imagery for detecting specific types of forest degradation, such as selective logging, and have investigated deforestation on a regional scale using the NDVI approach. Moreover, the use of Landsat MSS images has been explored for detecting deforestation, and radar images have been utilized for estimating forest loss. In the context of Horowpathana National Park in Sri Lanka, the use of Landsat satellite images and NDVI analysis can be employed to detect deforestation.

SECTION 3: Methodology

3.1. Study area and data collection

The investigations were performed as a Case study on Horowpathana National Park (2,570 km^2)) which is the 23rd national park of Sri Lanka. It lies between the north latitudes 8.6192525°N and east longitudes of 80.8301721°E E which is shown in the figure below.

The dataset used for this project is satellite images which are obtained from USGS Earth Explorer.

- Link - https://earthexplorer.usgs.gov/
- This data set includes details about LANDSAT 8 & 9 images with Metadata(text ,JSON , XML) and Band files(TIF,JPEG,).

3.2. Methodology

Approach

Satellite imagery plays a crucial role in various applications, including environmental monitoring. We aim to acquire high-resolution satellite imagery within a 2-kilometer radius of Horowpothana or specific coordinates defining a bounding box. 3

Preprocessing

Preprocess the satellite imagery data, including atmospheric correction, cloud removal, and image calibration. Calculate the vegetation index , such as the Normalized Difference Vegetation Index (NDVI), for each image. (using Landsat bands 4 (Red) and 5 (NIR))

Data Preparation

Organize the data into input-output pairs, where the input is the previous vegetation index, and the output is the calculated change in vegetation index (as a percentage).

Model Training

Design and train a Random forest model to predict the deforestation percentage,and ARIMA(Autoregressive Integrated Moving Average) Model to predict historical values , to learn the relationship between the previous vegetation index and the change in vegetation index. Utilize frameworks like TensorFlow or PyTorch for building and training the model. Detection Of Deforestation Using Satellite Images in Sri Lanka

Model Evaluation

Assess the performance of the regression model using appropriate evaluation metrics such as mean squared error (MSE)or mean absolute error (MAE). Use validation datasets to evaluate the model's ability to predict the change in vegetation index accurately.

Model Testing

Apply the trained regression model to the testing dataset, which includes new, unseen satellite imagery data. Predict the change in vegetation index for these images and convert the output into a percentage value. 3.8. Post-Processing and Analysis Perform any necessary post-processing steps, such as aggregating predictions or visualizing the change in vegetation index over the area of interest. Analyze the results to identify deforestation or afforestation trends

based on the predicted percentage change values.

Normalized Difference Vegetation Index (NDVI):

The Normalized Difference Vegetation Index (NDVI) is a measure of the amount and vigor of vegetation on the land surface and NDVI spatial composite images are developed to more easily distinguish green vegetation from bare soils. NDVI has been one of the most commonly used vegetation indices in remote sensing since its introduction in the 1970s. With the increased availability of remotely sensed imagery from satellites and UAVs, more and more people have come to adopt NDVI in their activity beyond the scope of science.

In general, NDVI values range from -1.0 to 1.0, with negative values indicating clouds and water, positive values near zero indicating bare soil, and higher positive values of NDVI ranging from sparse vegetation (0.1 - 0.5) to dense green vegetation (0.6 and above).

How To Calculate NDVI and How does it work

NDVI is calculated with the following expression:

$$NDVI = (NIR - Red) / (NIR + Red)$$

where NIR is near-infrared light and Red is visible red light.

The NDVI formula, we need to take the reflectance value in two bands: the visible red band and near-infrared band. We won't be able to calculate NDVI by using natural color imagery or another type of band composites, even though they may contain the required bands. where,

RED = the red portion of the electromagnetic spectrum (0.6-0.7 μm) and

NIR = the near infrared portion of the electromagnetic spectrum (0.75-1.5 μm).

ARIMA Model

ARIMA stands for Autoregressive Integrated Moving Average. It is a statistical model used for forecasting time series data. The ARIMA model is a recursive model, which means that it relies on past calculations to make predictions about the future.

The ARIMA model is made up of three components:

- The autoregressive (AR) component: This component models the relationship between the current value of the time series and its past values.
- The moving average (MA) component: This component models the relationship between the current value of the time series and the errors of past predictions.
- The integrated (I) component: This component is used to make the time series stationary, which means that its mean and variance are constant over time.

The ARIMA model is typically denoted as ARIMA (p, d, q), with each letter signifying one of the three components as described:

- 'p' signifies the quantity of autoregressive (AR) terms.
- 'd' designates the degree of differencing required.
- 'q' indicates the count of moving average (MA) terms.

Random Forest

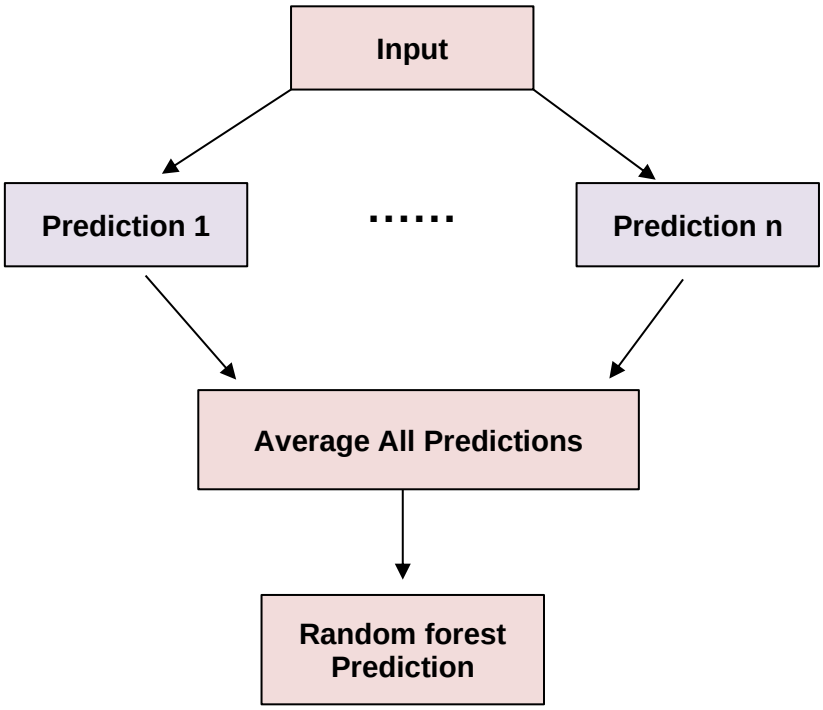
Random forest is a supervised machine learning algorithm that creates a model by constructing multiple decision trees. It is a versatile algorithm that can be used for both classification and regression problems.

This algorithm works by first creating a bootstrap sample of the training data. A bootstrap sample is a random sample of the training data with replacement. This means that some data points may be included in the bootstrap sample more than once, while other data points may not be included at all.Once the bootstrap sample has been created, the random forest algorithm builds a decision tree for each data point in the bootstrap sample. The decision tree is built by recursively splitting the data into smaller and smaller subsets until each subset contains only data points of the same class.This decision tree is built by using a greedy algorithm that chooses the best split at each node. The best split is the split that minimizes the impurity of the data. Impurity is a measure of how mixed the data is at a node.

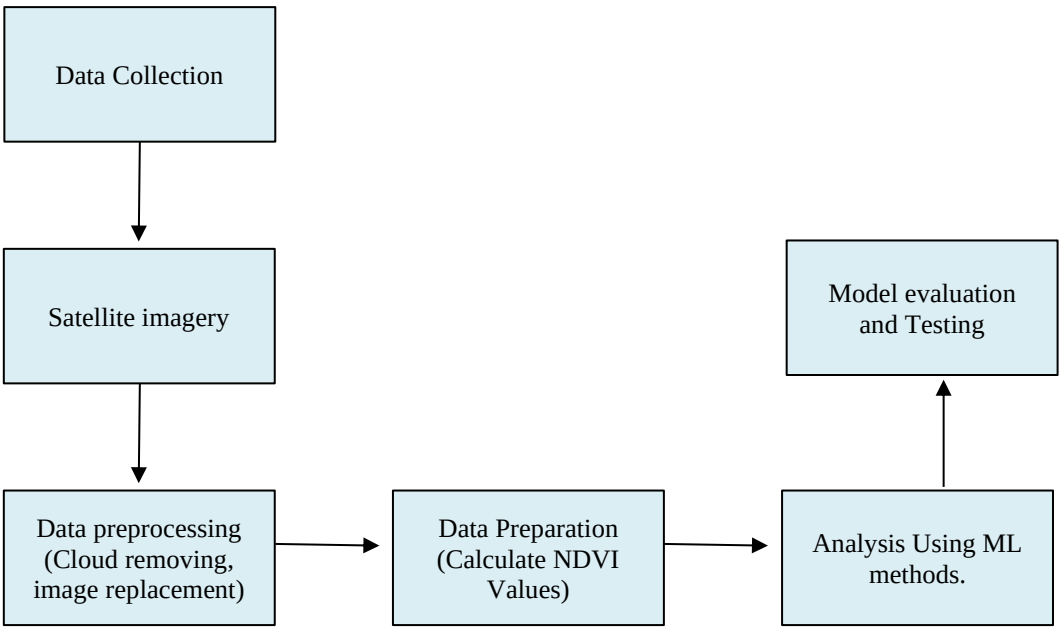
Random forest algorithm builds multiple decision trees by repeating the process of creating a bootstrap sample and building a decision tree for each data point in the bootstrap sample. The

number of decision trees that are built is a hyperparameter of the random forest algorithm.

It makes predictions by combining the predictions of the individual decision trees.Predictions of the individual decision trees are combined by taking the majority vote for classification problems or the average for regression problems.



Flowchart of methodology



Tools used.

For all the prepossessing, Analysing and other computational parts, we used “Jupyter Notebook”.

Satellite Image Sources

Analysis Platform

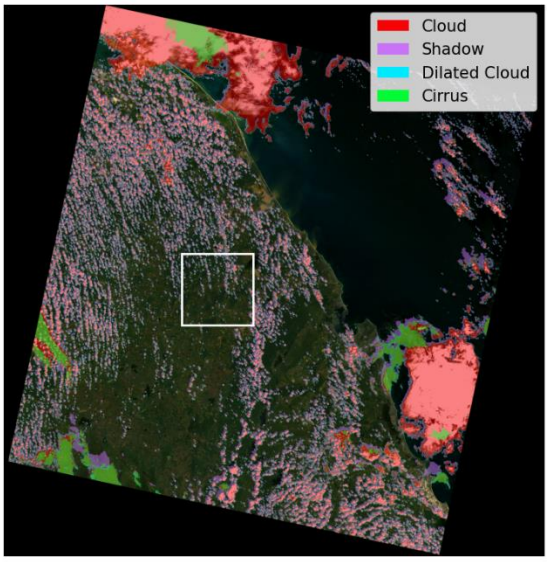
Python Libraries

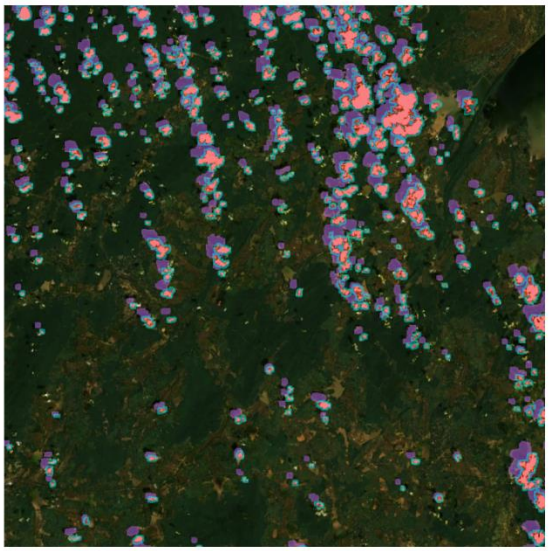
SECTION 4: RESULTS AND DISCUSSION

4.1. Results

Cloud imputation and Cropping images

Identify the different types of clouds in the satellite images. We observed 4 types of clouds such as Cloud Shadow Dilated Cloud and Cirrus.



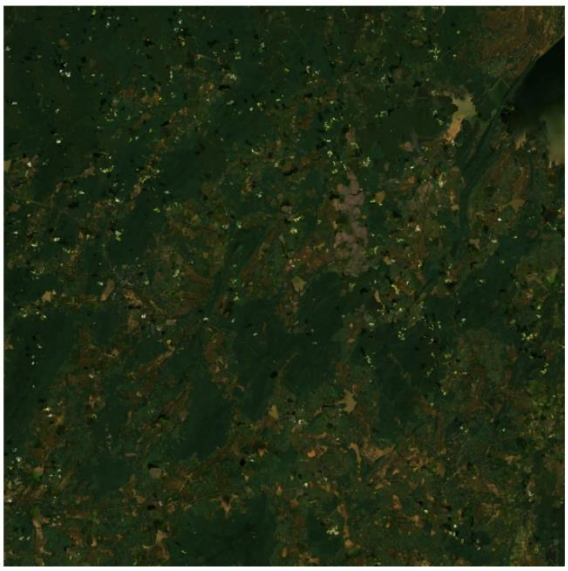


We used a satellite image with less clouds to imprint a satellite image with clouds.

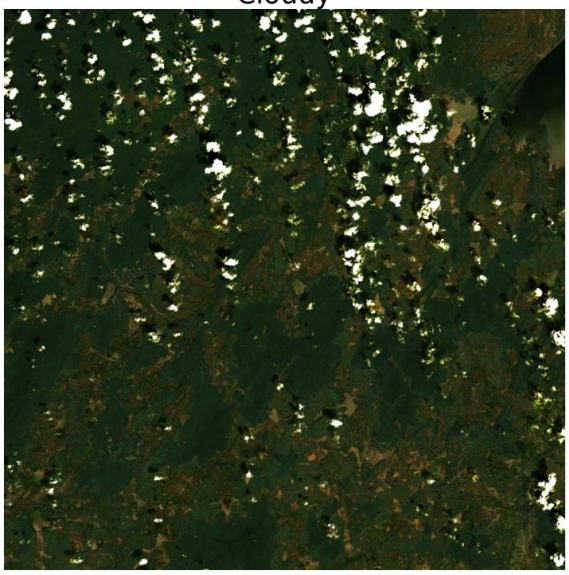
2020-06-14

2022-02-04

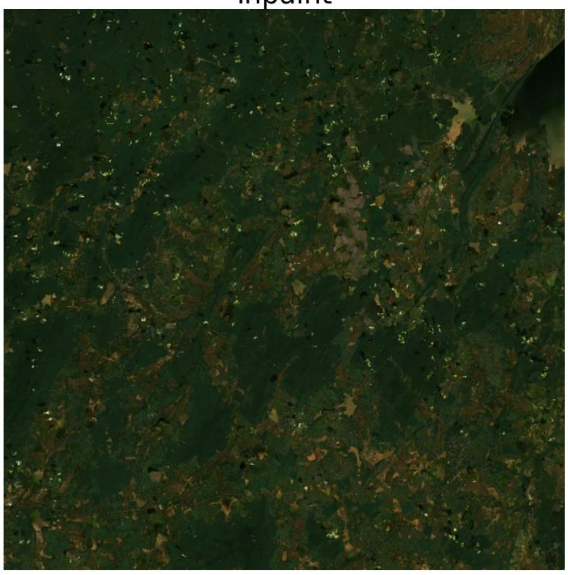
After removing the clouds, we get our area of interest which is Horowpathana National Park by cropping the removed image.



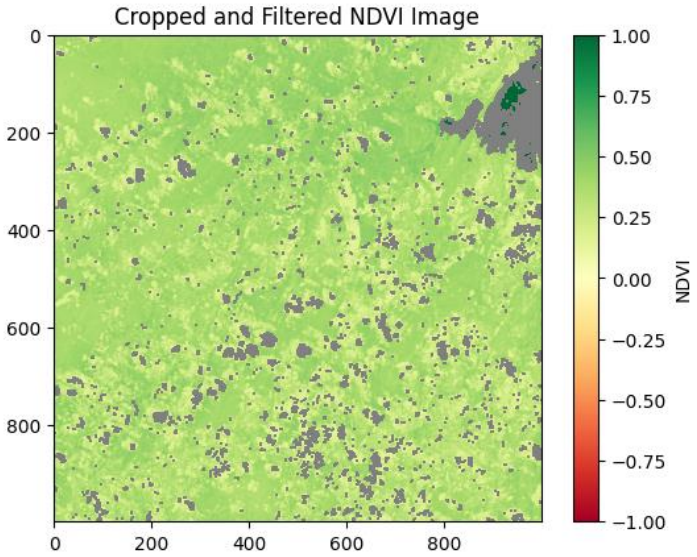
Cloudy



Inpaint



NDVI Calculation



Then we got an NDVI image based on its indices and then calculated the average ndvi values for each image. Then we used the ARIMA model to predict the percentage.

	Folder	Average_NDVI
2	LC08_L2SP_141054_20200614_20200824_02_T1_replaced	0.360736832
3	LC08_L2SP_141054_20200716_20200912_02_T1_replaced	0.285379829
4	LC08_L2SP_141054_20200817_20200920_02_T1_replaced	0.274821684
5	LC08_L2SP_141054_20200918_20201005_02_T1_replaced	0.26757831

Calculating deforestation Percentage

After that, we Calculated the deforestation percentage compared to the 2020/06/14 image.

	date	Average_NDVI	Deforestation_Percentage
2	6/14/2020	0.360736832	0
3	7/16/2020	0.285379829	20.88974447
4	8/17/2020	0.274821684	23.81657207
5	9/18/2020	0.26757831	25.82451061
6	10/4/2020	0.277102845	23.18421055

Model Evaluation

Mean Absolute Error (MAE): 0.29767006518816885

Mean Squared Error (MSE): 0.15406287452606163

R-squared (R2) Score: 0.9982258434435329

Mean Absolute Error (MAE) and Mean Squared Error (MSE) measure the average difference

and squared difference between predicted and actual values, respectively. A lower value for both indicates better model performance. R-squared (R2) Score measures the proportion of variance in the dependent variable explained by the independent variables, with a higher score indicating a better fit of the model to the data. In this case, the MAE and MSE values are relatively low, and the R2 score is very close to 1, indicating that the Random Forest (regression) model is performing very well and is likely to be accurate in its predictions.

4.2. Discussion

The purpose of this study was to examine how cloud cover affected the analysis of land use in Horowpathana National Park using satellite imagery. In order to improve accuracy, we used a cloud removal technique and a reference image without any clouds. These four distinct cloud types are Cloud Shadow, Dilated Cloud, Cirrus, and Cloud Shadow. We determined the Normalised Difference Vegetation Index (NDVI) to measure the amount of vegetation cover by isolating Horowpathana National Park in the cloud-cleared image.

An AutoRegressive Integrated Moving Average (ARIMA) model for time-series forecasting was applied to the NDVI values, which are crucial for evaluating changes in land cover. A remarkable R-squared (R2) score of 99.82% shows how accurately this method predicted future vegetation cover percentages. It was possible to determine a by comparing the vegetation cover to the reference image of 2020/06/14. Overall, model's robust performance, evidenced by low Mean Absolute Error (MAE) and Mean Squared Error (MSE) values of 0.30% and 0.15%, respectively, underscores its reliability in accurate land use analysis.

4.3. Limitation and Challenges

- Cloud Cover: Cloud cover can obstruct satellite imagery, making it challenging to acquire clear and continuous data, particularly in regions with frequent cloud cover.

- Data Availability: Access to high-quality, up-to-date satellite data can be costly and restricted, leading to potential gaps in coverage.

- Interpretation and Validation: Accurate interpretation of satellite data requires expertise, and on-ground validation can be logistically complex.

- Data Processing: Processing large volumes of satellite data can be computationally intensive, requiring suitable hardware and software.

4.4. Future work

During this research project we focussed on a Horowpathana National Park in Sri Lanka. Therefore, this project can be extended more covering larger area in future work. When we consider the technical side of the project our prediction model is based on the normalized difference vegetation index (NDVI). So this model can be improved more with different approach which uses different index to calculate forest density. Also, the accuracy of this model can be increase in several ways. One method is identifying the more factors that could affect the deforestation in the specific areas This means you can have more data. Hence having more data will affect in the quality of the training data as that will provide a broad picture of the problem.

SECTION 5: Conclusion

Through the application of ARIMA, we were able to analyse historical deforestation data and identify underlying trends and patterns.

Furthermore, the integration of Random Forest, a powerful machine learning algorithm, enabled us to explore the multifaceted factors contributing to deforestation. By considering NDVI values, we gained an understanding of the facts behind deforestation. This information is instrumental in developing targeted strategies for conservation and sustainable land management.

SECTION 6: References

Auhl, M. (2021, August 6). *What is an ARIMA Model?. Taking a quick peek into ARIMA modeling | by Miranda Auhl*. Towards Data Science. Retrieved August 21, 2023, from https://towardsdatascience.com/what-is-an-arima-model-9e200f06f9eb

Chiteculo, V., Abdollahnejad, A., Panagiotidis, D., Surový, P., & Sharma, R. P. (2018, 12 24). Defining Deforestation Patterns Using Satellite Images from 2000 and 2017: Assessment of Forest Management in Miombo Forests—A Case Study of Huambo Province in Angola. *Sustainability*, *11*(1), 98. 10.3390/su11010098

Liyanawatte, C., & Dias, K. (n.d.). *Analysis on Deforestation and Environmental Law in Sri Lanka*. IR@KDU Repository. Retrieved August 18, 2023, from http://ir.kdu.ac.lk/handle/345/1754

PRODES — Satellite Monitoring of Deforestation in the Brazilian Amazon Forest. (n.d.). OBT/INPE. Retrieved August 18, 2023, from http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes

Random Forest Regression Explained with Implementation in Python. (n.d.). Medium. Retrieved August 21, 2023, from https://medium.com/@theclickreader/random-forest-regression-explained-with-implementation-in-python-3dad88caf165

Serasinghe, B. (n.d.). *(PDF) GIS & Remote Sensing to Find Deforestation in Sri Lanka during the Period of 2000 to 2020 and Examining How to Build an algorithm to Predict Deforestation*. ResearchGate. Retrieved August 18, 2023, from https://www.researchgate.net/publication/342314218_GIS_Remote_Sensing_to_Find_Deforestation_in_Sri_Lanka_during_the_Period_of_2000_to_2020_and_Examining_How_to_Build_an_algorithm_to_Predict_Deforestation

Shimabukuro, Y. E., Arai, E., Duarte, V., & Anderson, J. (2019, 02 17). Monitoring deforestation and forest degradation using multi-temporal fraction images derived from Landsat sensor data in the Brazilian Amazon. *International Journal of Remote Sensing*. 10.1080/01431161.2019.1579943