Development and Comparison of Machine and Deep Learning Models for the Prediction of Land Degradation

Joshua Edwards

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science

University of North Carolina Wilmington

2022

Approved by

Advisory Committee

Dr. Karl Ricanek	Dr. Jeffrey Cummings
Dr. Narcisa Pricope	Dr. Gulustan Dogan Chair

Accepted by

Dean, Graduate School

CONTENTS

ABSTRACT

Background: The change of lush fertile and workable land into a state of desolation or unproductivity is known as land degradation. Land degradation, as a whole, is a serious problem; one which causes immeasurable damage every year. Despite the problem itself not being new the use of artificial intelligence to classify and predict land degradation is relatively new, especially compared to the problem itself. That being said numerous other researchers have done work using artificial intelligence to solve the problem of land degradation with varying results depending on the algorithm and dataset used.

Objectives: The primary purpose of this thesis is to develop and compare artificial intelligence algorithms to determine which gives the best predictions on variables related to land degradation. To this end, the objective is to train and compare both machine learning and deep learning models on the same data to determine what model will give the highest accuracy in accurately predicting land degradation variables.

Methods: Data for this thesis was taken from satellite imagery and readings from ground stations. Data used included precipitation, temperature, and ground cover (EVI) readings. For the predictions, different machine and deep learning models were developed including Random Forest, Gradient Boosting, and an LSTM. These algorithms were trained on approximately sixteen years of data and using those historic values would predict the expected EVI value for the following sixteen days.

Results: At the end of the study, after comparing both the machine and deep learning methods it was found that overall machine learning vastly outperformed the deep learning methods. In the end, random forest was the most accurate with a mean absolute percent error of 10.52%, and the top three models were all based on decision trees.

Conclusions: From the results gathered, it was concluded that, for the prediction of land degradation, machine learning-based decision tree models are the best choice. Out of these models, random forest was shown as the best from the results gathered, and overall

decision tree models outperform deep learning methods for this type of problem.

DEDICATION

IN PROGRESS

ACKNOWLEDGMENTS

IN PROGRESS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

1 Introduction

1.1 Motivation

Fertile land is one of the most important resources on the planet as it is essential for human development [1]. When land degrades human development suffers and lives are adversely affected. This can be seen in cases where degraded land has caused widespread famine such as it has in Ethiopia. Ethiopia is one of the most food-insecure countries in the world and this problem is exacerbated by land degradation with more than 85% of its land being degraded in some form [2][3][4].

Land degradation, however, is not limited to Africa; it affects over one billion people worldwide and has had a financial impact of over $10 trillion [5]. With such wide-spanning implications, accurate prediction of land degradation is essential to both saving money and improving lives. Therefore, this project focused on comparing artificial intelligence algorithms to see which ones provide the best accuracy for the prediction of land degradation.

Given the volatile nature of climate variables, coupled with the ongoing effects of climate change, accurate prediction is difficult. To make predictions, a plethora of variables are needed such as precipitation, temperature, and soil data. These along with many others are used when making predictions. The large number of variables coupled with their volatility from season to season can make prediction difficult. However, artificial intelligence excels in finding connections and relationships in data. Therefore, it is the perfect medium to use for prediction.

Currently, many researchers are engaged in using artificial intelligence to predict land degradation. Their methods include different algorithms including Random Forest (RF), classification and regression tree (CART), and support vector machine (SVM) [6][7][8][9]. Similarly, datasets vary widely between different approaches and areas. Some studies used data that was collected in person for the area they were studying [6][8][10]. Others used data that was collected remotely through weather stations or satellites [11][7].

1.2 Objectives

The first goal of this thesis is to train and test a variety of artificial intelligence (AI) algorithms to predict an average land cover (EVI) value for a sixteen day time period. The second and primary goal of this thesis is to then compare these algorithms and determine which one will give the highest accuracy for this predicting land degradation variables.

1.3 Research Question

Based on previous research we know that land degradation can be predicted with a relatively high degree of accuracy using artificial intelligence[6]. The questions this thesis answered are as follows:

1. What algorithm performs the best for the prediction of land degradation?
2. What type of AI algorithm will perform better machine or deep learning?

To answer these questions, a variety of different machine learning and deep learning algorithms were trained on the same dataset, and they were then tuned to determine which one and what method worked best on the environmental data.

1.4 Research Novelty

The novelty of this research comes from doing a comparison of machine and deep learning to gauge which is best for the prediction land degradation. Other related works tend to focus mainly on predicting land degradation by using artificial intelligence. They do sometimes use multiple algorithms and compare the results, but their focus is usually not on the algorithms themselves. The main focus of this paper is what type of algorithms will, in general, perform best for the prediction of land degradation.

1.5 Research Scope

For this study, multiple algorithms were applied to the dataset which is elaborated upon in the data exploration section. The data is limited to the country of Columbia; however,

the algorithms and methods applied in this thesis could be applied to other areas if data is pulled from the same data sources.

1.6 Thesis Organization

This thesis has been split up into the following sections Introduction, Background, Related Work, Methods, Results, and Discussion. In the introduction, the motivation of why this research was conducted was explained, and the specific research questions that were being asked are listed. The scope of the problem is also listed there. In the Background section, the history of both the topic of land degradation and the background of the algorithms used in this study are talked about. In the following section, titled Related Work, similar works related to this project are talked about. The Methods section lists information about the data that was used to get the final results. It also goes into detail about how the data was explored to find interesting observations and how the data was cleaned to achieve the final dataset. It also goes into what was done to obtain the final results including the algorithms used and the details of their construction. Finally, Discussion details the results achieved with an estimation of why those results were obtained. It also discusses some future applications of this work and wraps everything up with a conclusion.

2 Background

2.1 Land Degradation

Lush fertile livable ground is perhaps one of, if not the most important resource that humanity has. Without it, the ability to grow crops for the expansion of humans as a species would be drastically reduced to a point where mankind would struggle to exist. This is why accurate forecasting and prediction of land degradation is important, to preserve the fertile lands that we currently rely upon to survive.

While there is some disagreement over the exact definition of what the phrase land degradation encompasses it is generally accepted to mean, "The temporary or permanent decline in the productive capacity of the land, and the diminution of the productive potential, including its major land uses (e.g., rain-fed arable, irrigation, forests), its farming systems (e.g., smallholder subsistence), and its value as an economic resource" [12]. Land degradation itself is not a new concept or new problem by any means. Some of the earliest known examples are millennia-old [13]. One of these comes from ancient Mesopotamia around the year 2400 BCE where human irrigation efforts led to the salinization of lands which caused them to become commercially unproductive for growing crops and therefore degraded [13].

Since that time research has been plentiful with thousands of papers being published on the topic, not all of them recent. An article titled "African Survey" published in 1938 called it the "Scourge of Africa" [12][14]. This continued interest in land degradation throughout the years is for good reason, as recent estimates show that, currently, land degradation affects about 25% of the world's surface [15]. If that alone is not concerning enough, if current trends were to continue unchanged that number would rise to 95% by 2050 [15].

With the continued interest in land degradation throughout the years, it should come as no surprise that it is by far a complex issue with many different causes. However, for this study, two variables were chosen to act as predictors of land degradation precipitation

and temperature. Studies have found a strong correlation between precipitation rates and rates of land degradation [16]. Similarly, temperature has also been shown to be one of the driving metrics of land degradation [16]. It is for this reason that these two metrics were chosen as the predictors for this study.

2.2 Machine and Deep Learning Algorithms

In this study, several machine and deep learning methods were explored to see which ones provided the best results. Below is an explanation of each of the algorithms that were used. They are split up into both machine and deep learning sections.

2.2.1 Machine Learning

Machine Learning is one of the branches of artificial intelligence and also one of the oldest. The term "machine learning" was first coined in 1959 [17]. An apt description of machine learning is given here, "Computer systems perform functions through (Machine Learning) such as clustering, calculations, and pattern identification. The learning process is attained using various algorithms and arithmetic structures to analyze the information. This information is classified by some characteristics called features. (Machine Learning) is used to find a relationship between the features and some output values called labels" [18]. With that being said numerous different algorithms fall under the proverbial umbrella of machine learning. Those used in this study are included below.

Random Forest: Random forest is an ensemble supervised learning method that uses a combination of many decision trees to make its predictions [19]. In summary, it uses multiple decision trees and averages the output of the trees to make a strong prediction [19]. Random forest is commonly used in land degradation prediction as can be seen from some of the related works. This is the main reason why it was included in this study as the baseline model.

K-nearest Neighbors (KNN): KNN is a supervised learning technique and one of

the most commonly used machine learning algorithms [20]. KNNs work by predicting the output variable as an average of nearest observations [21]. KNNS also use, "A set of k-nearest observations to decide on the response value of a test case thus trying to minimize the effect of outliers in a training dataset" [20]. KNN's ability to sometimes beat other more complex methods and its ease of implementation is why it was included in this study [20].

Linear Regression: Linear regression is perhaps the simplest and most common machine learning model [22]. Linear regression is a supervised learning method that functions by simply trying to get a linear fit between the dependent and independent variables [22]. Its use in this study is primarily to see if there is a somewhat linear relationship between the input and output variables of the dataset. It is also used as a baseline model because any model that performs worse than a linear fit is not worth pursuing any further.

Gradient Boosting Regression: Gradient boosting works by, "Consecutively [fitting] new models to provide a more accurate estimate of the response variable" [23]. Natekin and Knoll also aptly described the idea behind gradient boosting by saying, "The principle idea behind this algorithm is to construct the new base learners to be maximally correlated with the negative gradient of the loss function" [23]. Gradient boosters also have a high degree of customization that makes them perfect for several different applications [23]. The ability to be adapted to a wide variety of problems is one of the main reasons why it was chosen as one of the test algorithms for this project.

Extreme Gradient Boosting Regression (XGBoost): Extreme gradient boosting is very similar to normal gradient boosting; however, there are slight differences that can lead to an increase in performance. Xgboost reduces the run time compared to regular gradient boosting [24]. It also uses a different regularization than standard gradient boosting which helps improve model performance [25]. This improvement over standard gradient boosting along with the fact that in some cases xgboost outperforms random forest is why it is included in this study.

Support Vector Machine: Support vector regression is a supervised learning algorithm that essentially creates a hyperlane (or in simpler terms line) which then splits the data into multiple groups [26]. SVMs have been used in the past with some success for the problem of time series prediction [26]. Due to its success in the past with time series prediction, it was chosen for this study.

Decision Tree: Decisions trees are a supervised learning technique that work by, "Recursively partitioning a data set and fitting a simple model to each partition" [27]. Ultimately they make predictions based on how problems were solved previously. Decision trees are widely used due to how easy they are to construct and their generally high degree of reliability [28]. This easy setup and ability to usually produce good results is part of why decision trees were tested in this study.

K Means: K means differs from the other machine learning models used in this study as it is an unsupervised technique [29]. K means essentially functions by clustering together k samples and calculating the distance to all of the other samples in the dataset until eventually splitting them into k groups repeatedly [29]. K means was chosen for this study just to test how an unsupervised algorithm would perform on the data compared to the previously mentioned supervised algorithms.

2.2.2 Deep Learning

Deep learning is the newest subset under the AI umbrella with its origins starting in the 1980s and the term itself arising in 2006 [30]. Deep learning is summarised as an AI algorithm that seeks to emulate the function of a human brain through the use of neural networks [30]. While deep learning algorithms have a vast array of applications and uses they tend to suffer from being far more complex than machine learning algorithms [30]. That being said some problems tend to be more easily solved with machine learning over deep learning or vice versa. For this study, Three algorithms were used to get a good sense of how deep learning compares to the more commonly used machine learning, for land

degradation prediction.

Feed Forward Network: Feed-forward neural networks are one of the two main network architectures when talking about artificial neural networks [31] Essentially, a feed-forward network means that there is, "no “feedback” from the outputs of the neurons towards the inputs throughout the network" [31]. The feed-forward network used in this thesis was a simple model used mainly for validation of concept purposes.

Long Short Term Memory (LSTM): LSTMs are recurrent neural networks that use gates to regulate the flow of information through the network [32]. These gates determine what information gets added or subtracted as the data makes its way through the network's cells [32]. Ultimately LSTMs seek to solve the problem of vanishing gradients that plague recurrent neural networks because of this LSTMs excel at time series prediction[32]. It is for this reason that an LSTM was chosen to be the main algorithm of the deep learning section of this thesis.

Multilayer Perceptron (MLP): Multilayer perceptrons are the most commonly used neural networks [33]. MLPs are a type of feed-forward network and consist of input, hidden, and output layers [33]. Data moves through the layers and an activation function before being backpropagated through the network until the values have converged [33]. Since they are commonly used it was decided to use one as a baseline for the deep learning section of this thesis.

3 Related Work

For this topic, several similar works will be summarized below. In addition, they are also summarized more concisely in Table 1. Yousefi et al.'s work sought to assess the extent of land degradation in Iran [6]. Their primary purpose was to determine patterns behind land degradation and to formulate strategies for the management of land and resources. The study used data derived from measurements taken manually from fields in Iran. These measurements made up the entirety of the dataset. This study implemented three different models to make predictions. These models were random forest, support vector machine, and classification and regression tree. The study found that out of these three models the most accurate was random forest with an overall accuracy of 96%.

Lead Author	Description	Dataset	Methods	Accuracy
Yousefi et al. [6]	Examined land degradation in Iran by comparing three different algorithms.	Tabular data. Taken from farm measurements Sample size = 1147	Random Forest Support Vector Machine Classification and Regression Tree	96%
Rukhovich et al. [11]	Examined land degradation in Russia using convolutional neural networks.	Visual Data. Taken from the Landsat dataset Sample Size = 544,840	Convolutional Neural Network	87.5%
Nzuza et al. [7]	Examined land degradation In South Africa using random forest regression.	Visual Data. Taken from CHIRPS dataset and Sentinel-2 dataset Sample Size = 36 field plots of size 20m x 20m	Random Forest	92%
Vagen et al. [8]	Examined land degradation in Ethiopia using random forest and gradient boosting.	Visual and Tabular Data, Taken from the Landsat dataset and field samples Sample Size = 38 sites each 100 square km	Random Forest	80%
Torabi et al. [9]	Examined land degradation in Iran using support vector machines.	Tabular Data Taken from numerous sources Sample Size = 400 sites	Support Vector Machine	88%
Cerretelli et al. [10]	Examined land degradation in Ethiopia using Linear Regression	Visual and Tabular Data Taken from various datasets and field samples	Linear Regression	89%
Pal et al. [34]	Investigated the contributing factors of land degradation in India	Tabular Data Taken from field samples	Boosted Regression Tree	93%
Yacine et al. [35]	Developed a model to predict landslides due to land degradation in Algeria.	Tabular Data	Random Forest Gradient Boosting Regressor	90%
Habibi et al. [36]	Used artificial intelligence to predict ground water level as a cause of land degradation in Iran.	Tabular Data	Artificial Neural Network	96%
Chakrabortty et al. [37]	Used artificial intelligence to predict soil erosion in India.	Image Data Taken from satellite imagery	Analytical Neural Network and Geographically Weighted Regression	91%
Garg et al. [38]	Created a simple artificial neural network to predict soil erosion.	Tabular Data Taken from another paper	Artificial Neural Network	94%
Abolhasani et al. [39]	Developed a new framework to predict land degradation.	Visual Data Taken from satellite imagery	Random Forest	81%

Table 1: Summary of Related Works

D. I. Rukhovich et al. focused on classifying land degradation in Russia [11]. The primary goal of the paper was to prove the validity of using remotely sensed data to classify and predict land degradation. This paper used visual data taken from the Landsat dataset. Their primary method involved using the normalized difference vegetation index (NDVI) to determine which areas of land had decreased amounts of biomass. The sample area for this study consisted of fields split into plots. The total size of the dataset was 544,840. The study used a convolutional neural network model to make a binary classification as to what areas were degraded. The overall accuracy was 87.5%.

Nzuza et al. focused on land degradation in South Africa [7]. Their primary motivation for the research was based on the need for real-time monitoring of land degradation. This paper used visual data taken from visual datasets including the CHIRPS dataset along with manually collected field samples. This study focused on monitoring land degradation conditions and classifying which areas were at an increased risk. The method used was random forest, and the overall accuracy was 92%.

Vagen et al. focused on land degradation in Ethiopia [8]. The primary reason for this paper's research was due to the widespread land degradation in Ethiopia. Proper identification of land degradation was the desired goal so that interventions could be made. The dataset used in this study was a combination of both visual and tabular data. This data came from field samples collected and the Landsat dataset. Landsat images were taken for 38 sites of approximately 100 square kilometers each. The model used was random forest, and the total accuracy was 80%.

Torabi et al. focused on the research goal to, "develop a new quantitative (land degradation) mapping approach using machine learning techniques, benchmark models, and human-induced and socio-environmental variables" [9]. This paper also focused on accurately mapping land degradation to prioritize land and water conservation efforts. This study focused solely on tabular data collected from field surveys. The study used four methods in testing but eventually found that overall SVM performed the best with an accuracy of 88%.

Cerretelli et al. focused on Ethiopia [10]. The goal of their study was to, "infer land degradation through (ecosystem services) assessment and compare the modelling results obtained using different sets of data" [10]. This paper used data taken from both MODIS and Sentinel datasets along with data taken from manual surveys. They used a simple linear regression model that gave them 89% accuracy.

Pal et al. focused on an area in eastern India for there study into land degradation [34]. The focus of their research was to, "investigate chemical weathering, gully erosion and cohesiveness through field-based measurements with a view to understand the controlling factors of potential land degradation" [34]. For their research, they used a combination of mostly decision tree algorithms which can be seen in Table 1, and their data was manually collected. Overall, this study had fairly accurate results. Their prediction metric of choice was receiver operating characteristic which out of the algorithms tested returned 0.93 for a boosted regression tree algorithm.

Yacine et al. sought to develop a model to accurately predict the susceptibility of areas to landslides due to land degradation [35]. The ultimate goal was to, "reduce the physical degradation caused by landslides and, to inspect what is required to properly control it" [35]. Their region of focus for the study was northeastern Algeria and they chose random forest and extreme gradient boosting to make their predictions. The accuracy for their study ended up being around 90% for both models used.

Habibi et al. focused on using artificial intelligence to predict land degradation by predicting ground water levels in the Sharifabad watershed region in Iran [36]. They tried various different models among them being Partial Least Square Regression, Artificial Neural Networks and Adaptive, Neuro-Fuzzy Inference System [36]. Overall their best performing model was an artificial neural network with an r-squared value of 0.963 and a mean squared error of 7.12

Chakrabortty et al. similarly to the previous study by Pal et al. studied an area in eastern India [37][34]. For this study, they sought to use artificial intelligence for the pre-

diction of soil erosion. They described their main goal was to, "identify areas vulnerable to soil erosion and propose the most suitable model for soil erosion susceptibility in sub-tropical environment" [37]. During their study they used an analytical neural network and geographically weighted regression ensemble method [37]. Their study found that they were able to achieve a final precision of 91.64 with their model. Their data was taken from the Sentinel 2 MSI and Landsat 8 OLI satellites.

Garg et al. sought to predict soil erosion by using an artificial neural network [38]. Their stated research purpose was, "This study aims to develop a simple Artificial Neural Network (ANN) based model to predict erosion of biochar amended soils (BAS) under varying conditions (slope length, slope gradient, rainfall rate, degree of compaction (DoC), and percentage of biochar amendments)" [38].Their data was taken from another paper's results which were taken manually. The results from their artificial neural network showed that its r-squared value was 0.939 for their top performing model.

Abolhasani et al. tried to develop a new framework for the modeling of land degradation [39]. They described their research goal as, "This research aimed to develop a new conceptual framework to predict LD susceptibility based on net primary production (NPP) and machine learning approaches" [39]. The data taken for this study came from the MODIS satellite. Their results showed that overall random forest provided the best accuracy for their problem with an AUC of 0.81 [39].

4 Data

4.1 Datasets

The data for this project consists of rainfall, temperature, and vegetation data. This data was acquired from Google Earth Engine and the IRI/LDEO Climate Data Library at Columbia University. Google Earth Engine is an online data repository and platform containing a large amount of satellite and geospatial datasets [40]. It is primarily focused on providing researchers easy access to resources for climate research. The IRI/LDEO Climate Data Library is an online data warehouse that similarly to Google Earth Engine contains vast quantities of data related to climate research [41].

The rainfall data for this project was gathered in the form of the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) dataset. CHIRPS consists of satellite imagery taken in approximately 40 years and measures rainfall through the use of both satellite and weather station data [42]. Similarly, the Climate Hazards Group InfraRed Temperature with Station data (CHIRTS) dataset functions in the same manner as CHIRPS; however, it contains temperature instead of rainfall data [43]. The final dataset for this project is NASA's MODIS Vegetation Index which uses satellite imagery collected from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite to determine the vegetation index value for a given area [44]. It does this by applying a mathematical formula $G * ((NIR - Red)/(NIR + C1 * Red - C2 * Blue + L))$ which uses the red, near-infrared, and blue wavelengths from the satellite to determine the degree of vegetation that an area possesses [45].

4.2 Data Cleaning

After compiling the data, the first step was to resize the datasets to the same resolution. Both CHIRPS and CHIRTS are sized at 5km x 5km resolution; however, MODIS operates at a 1km x 1km resolution. So, the MODIS dataset was resized using the reproject and

Metric	CHIRPS	CHIRTS	EVI
Years of Data	1981-2022	1983-2016	2000-2022
Frequency	Daily	Daily	16 Day Product
Date Range Used	2000-2016	2000-2016	2000-2016
Mean	29.90 (C)	7.08 (mm)	0.42
Median	30.68 (C)	0 (mm)	0.44
Min	4.09 (C)	0 (mm)	∼0
Max	40.87 (C)	523.51 (mm)	0.80
Std	3.85	12.35	0.11
Citation	[42]	[43]	[44]

Table 2: Measurements gathered from one column out of the dataset. Values are representative of the whole dataset with a slight margin of error.

resize functions from the Google Earth Engine Python library. Once this data was resized and reprojected each image was then converted into a tabular form of data. The conversion of data was accomplished by iterating through each pixel in a given image and placing the numerical value of the pixel into a cell in a .csv file. This was done for every image in all three datasets and the data was grouped by date and pixel value.

Once the data was put into the .csv file a problem arose with the EVI data being a 16-day product whereas the temperature and precipitation data was a daily product. To rectify this issue, the data was reshaped with each row consisting of 16 days of both temperature and precipitation data and two cells of EVI data. The two cells of EVI data represent the previous EVI value and the new EVI value. An example of the structure of the final data file can be seen in Table 3.

4.3 Data Exploration

While purely numerical data was more than adequate for the models used, it is unfortunately hard to visualize for humans. Therefore, it was decided that the data needed to be put into a visual format, to be easily interpreted. However, due to the large amount of data used in this project it was difficult to settle upon a suitable way to visualize the data in a way that would be both useful and easy to interpret.

Out of the options presented the best one was to take the average values of the different seasons, in Columbia, and track the changes in value from one season to another. This proved difficult as Columbia's position on the Equator means that there is no uniform seasonality throughout the entire country. In fact, throughout the year the temperature in Columbia does not vary drastically and remains relatively uniform. This can be seen in Figure 3. The only noticeable effects of seasonality in Columbia come from the wet and dry seasons. Different areas in Columbia experience seasons differently with some areas experiencing only one wet and dry season and others experiencing two of each; however, for the sake of research purposes, it is commonplace to generalize the whole of Columbia into two wet and dry seasons [46][47]. For this generalization, the data is generally split into three-month increments of December-February, March-May, June-August, and September-November [47]. An Example of this can be seen in Figure 1 where a change in vegetation value can be seen throughout the different seasons.

Visualizing a change in values for one year was helpful to gauge whether there was any significant change in values that could be predicted. However, to see any long-term change it was necessary to graph the change from year to year. To do this the entire sixteen years' worth of data were split into images similar to Figure 1 and then concatenated into a single graph of the entire dataset. This was done for not only the vegetation data but also for the precipitation and temperature data. This can be seen in Figures 2, 3, and 4

From these graphs, several important observations were made about the data. One observation made clear by looking at the graph of temperature in Figure 3 is that the average temperature does not change drastically from one season to the next. While there are some noticeable changes from season to season they are largely small compared to the change in precipitation and temperature data in Figures 2 and 4. This is not too surprising given Columbia's position close to the Equator. However, since temperature is an important predictor of land degradation it was still included as even the small changes in value could still help the model make its predictions.

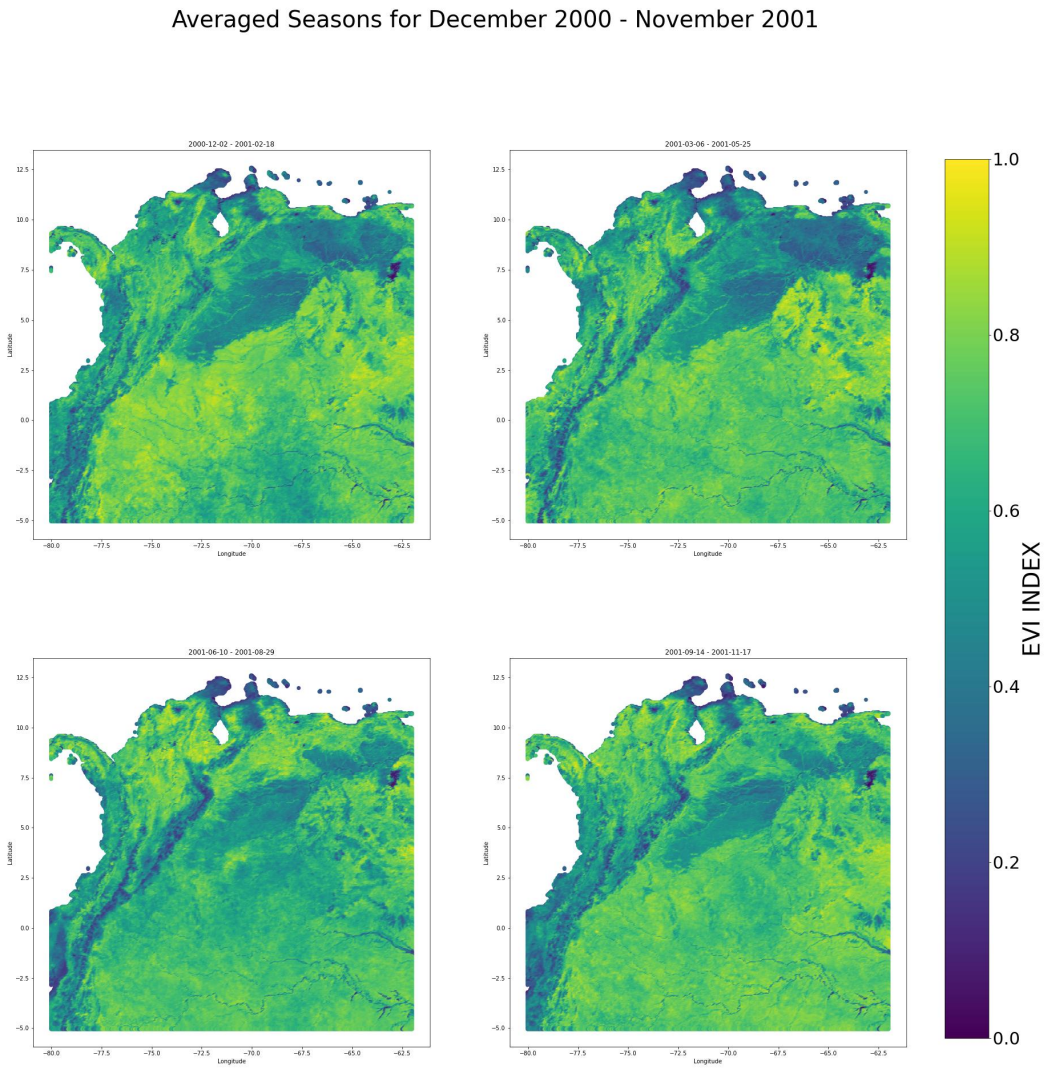

Figure 1: Seasonal averages of EVI for a period of December 2000 - November 2001

Looking at the figures it is also possible to notice trends in values from season to season and to notice changes in different variables simultaneously. For example, looking at periods of extremely heavy rainfall it can be seen that EVI goes down when precipitation is excessive; however, periods with moderate rainfall tend to show an increase in EVI values. This could be caused by excessive rainfall leading to an increase in mud and erosion and therefore a decrease in visible vegetation. Whereas moderate rainfall doesn't disturb the

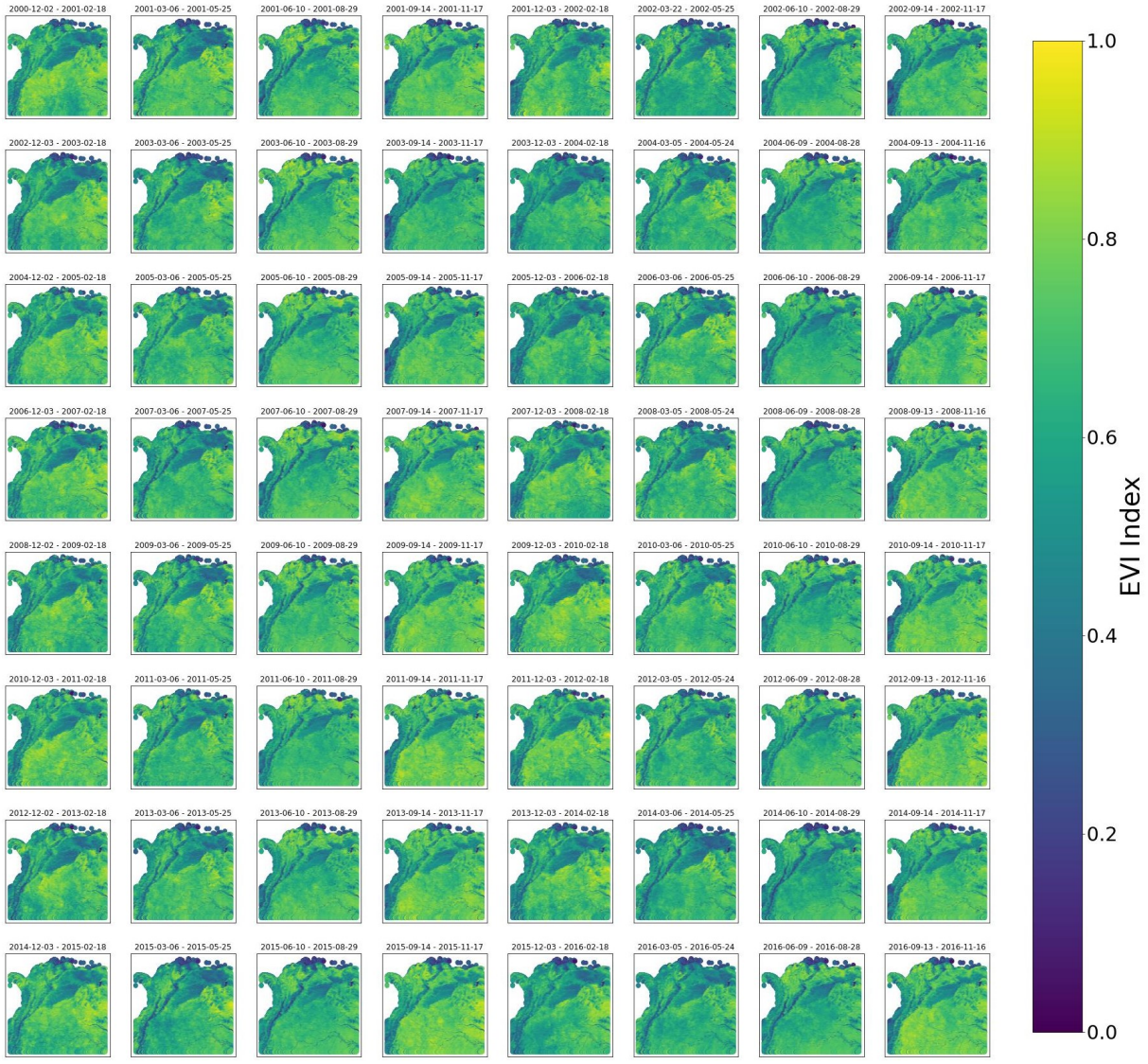

Figure 2: Seasonal averages of EVI for a period of December 2000 - November 2016

soil and shows more benefit to vegetation growth. It can also be seen that areas along the western coast that maintain a consistent level of rainfall from season to season don't show much change in EVI values. Overall, this shows positive signs that precipitation is a positive value for the models to use for predicting EVI values.

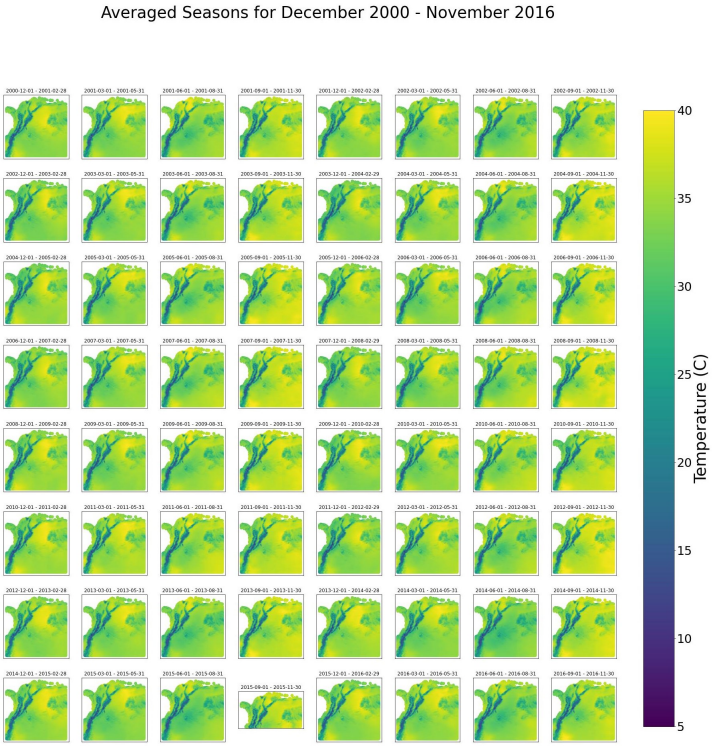

Figure 3: Seasonal averages of temperature for a period of December 2000 - November 2016 Note: The graph for the time frame of 2015-09-01 - 2015-11-30 is cropped differently than the rest due to missing data.

5 Experiments

For this project, to determine the best algorithm to use for the prediction of land degradation, several commonly used AI algorithms were developed and tested to see which ones provided the best accuracy. The below sections discuss what algorithms were used and how they were configured. The final results for each algorithm are compiled into tables with a discussion of the results included in the results section. For the construction of the algorithms, the base models were taken from the scikit-learn, Keras, TensorFlow, and XGBoost Python libraries [48][49][50][51].

Pixel	Day	Month	Year	Precipitation (t-15 ... t-0)	Temperature (t-15 ... t-0)	Previous EVI	New EVI

Table 3: Structure of the input dataset. Note for precipitation and temperature they both have 16 rows of data, and t represents the prediction date.

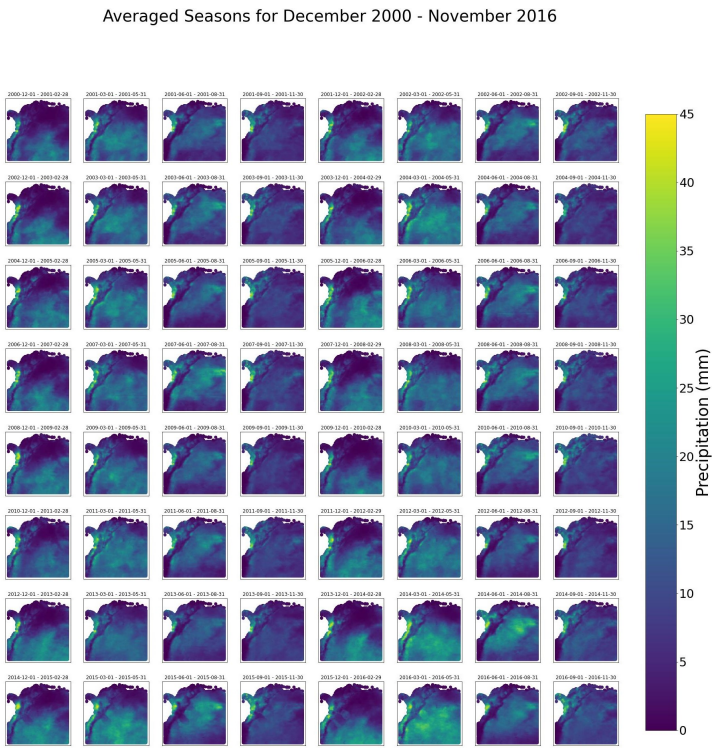

Figure 4: Seasonal averages of precipitation for a period of December 2000 - November 2016

The data for these experiments was split up as shown in Table 3. This data consisted of 38 columns and 29374848 rows of data. The original experiment involved training on the entire dataset other than a single row of data and then predicting on that final unused row. However, after several experiments, it was found that the algorithms' prediction accuracy varied drastically depending on the date being predicted. Some dates were more consistent than others so to get a consistent accuracy for the models the datasets were instead split using an 80-20 train test split to make predictions. The first 37 columns were the variables used to obtain the prediction and the 38th column was the predicted output variable.

5.1 Machine Learning

Based on the literature review it was found that a large number of studies that seek to predict land degradation or similar phenomenon tend to use mainly machine learning algorithms. For this reason, several different machine learning algorithms were tried and

the results were compared to see which ones performed best for this type of problem. Each algorithm developed and the configuration used are detailed in the below sections.

Random Forest: Random forest was chosen due to its common use for solving similar types of problems as demonstrated in the related works section. The input model used was a standard scikit-learn model with the number of n_estimators set to 32 where n_estimators is the number of trees in the forest. Due to the size of the dataset, the default value of n_estimators equals 100 proved to be too memory intensive to run on the machine used for this study. each time the number of trees was increased the accuracy of the model showed a slight degree of improvement, however, the number 32 was settled on to prevent memory errors.

K - Nearest Neighbors (KNN): KNN was chosen for this study as it is one of the easiest machine learning algorithms to understand and implement. However, the problem encountered was because KNN algorithms tend to struggle with large sets of data [52]. Due to the vast amount of input data from this study the KNN algorithm failed to reach an end state after over a week of calculations and was terminated.

Linear Regressor: Linear regression was used because it is a good algorithm to get a baseline prediction with as it is one of if not the simplest algorithm for machine learning. As such it wasn't expected that this would provide good results but instead, it was a good indicator of what algorithms were especially unsuited to this type of problem. For that reason, the algorithm used was just the standard scikit-learn implementation of a linear regressor.

Gradient Boosting Regressor: A gradient boosting regressor was chosen in part because of its relation to random forest regressors. Both are ensemble methods that use decision trees to make their predictions. For this reason, it would most likely have a similar degree of accuracy to that of random forest. For the implementation, the default gradient boosting algorithm from the scikit-learn library was used. While the accuracy for the standard model wasn't bad it still vastly underperformed random forest, so it wasn't

fine-tuned.

Extreme Gradient Boosting Regressor: Even though normal gradient boosting underperformed an extreme gradient boosting algorithm from the XGBoost Library was developed and tested. For the algorithm parameters, a max_depth of 20 and eta of 0.6 were passed in which means that the maximum depth for the trees was 20 and the learning rate was 0.6. The learning rate was experimented with between values of 0.1 - 1 and it was found that 0.6 gave the best degree of accuracy. The accuracy got better as max_depth increased but eventually memory usage became an issue so it had to be capped at 20.

Support Vector Machine (SVM): The next algorithm used was an SVM which was imported from scikit-learn. SVM is a commonly used algorithm used for a variety of different problems; therefore, it was decided to import and train a default implementation from scikit-learn to gauge how it performed on this type of problem. However, the initial test run underperformed linear regression so the model was not improved upon.

Decision Tree: Running a decision tree algorithm seemed to be a good choice for getting another good baseline estimation of accuracy for the models. Since random forest uses multiple decision trees to make its predictions running a decision tree algorithm for comparison seemed appropriate. For this implementation, a default decision tree from scikit-learn was trained on the data without any changes to the default arguments.

K Means: The final machine learning algorithm test was a K means regressor. Since K means is an unsupervised learning technique it was different than the rest of the methods tried. While it was assumed that it would not perform nearly as well as a supervised learning technique the assumption was tested anyway to make sure. A default model from scikit-learn was imported and trained. This confirmed previous assumptions as the model vastly unperformed all of the supervised methods used. Therefore, there was no reason to fine-tune the model.

5.2 Deep Learning

One of the main questions of this project was whether deep learning could provide a higher level of accuracy compared to the more standard machine learning methods. Based on past literature it is already known that machine learning is sufficient to make predictions for land degradation; however, considerably fewer researchers have used deep learning to make their predictions. So, one of the main questions to be answered was whether or not these common deep learning algorithms could outperform traditional machine learning methods.

Feed Forward Network: The first model that was developed and trained was a simple feed-forward network with an architecture as shown in Figure 5. This was less of a serious test and more of a way to test the data on a model and make sure the setup was correct before developing a more complex model. That being said this model was run for a single epoch; however, it performed worse than most of the machine learning algorithms and wasn't fine-tuned any further.

Long Short Term Memory (LSTM): To make a prediction on the input data, it was hypothesized that an LSTM model would provide a good prediction. Since LSTMs are commonly used for time series prediction and since they are good for finding patterns in long-term data it seemed like the best bet for making predictions on the input data. The data passed to the LSTM was the same as the rest of the algorithms and the architecture of the model is shown in Figure 6 Multiple different versions of the architecture were tested but the architecture shown in Figure 6 showed the highest accuracy out of any that tested.

Multi Layer Perceptron (MLP): The final deep learning method tested was an MLP. MLPs are good at learning from non-linear data and they are also commonly used for extremely complex problems. Therefore, based on the struggle of getting an accurate prediction an MLP was tested to see if the baseline results would be more accurate than the LSTM model. For the implementation, the default MLP was imported from scikit-learn

and trained with the input data to see if the baseline results were more accurate than the LSTM model. However, the base model was less accurate so it wasn't fine-tuned anymore.

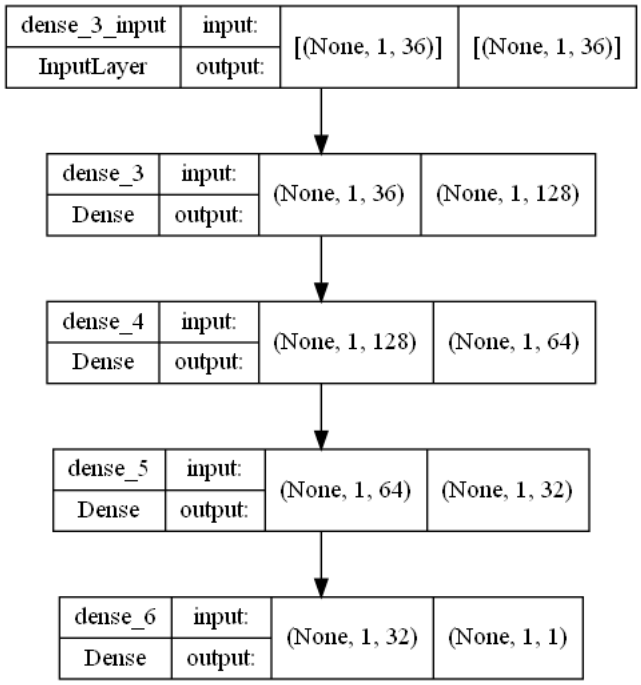

Figure 5: Feed Forward Model Architecture

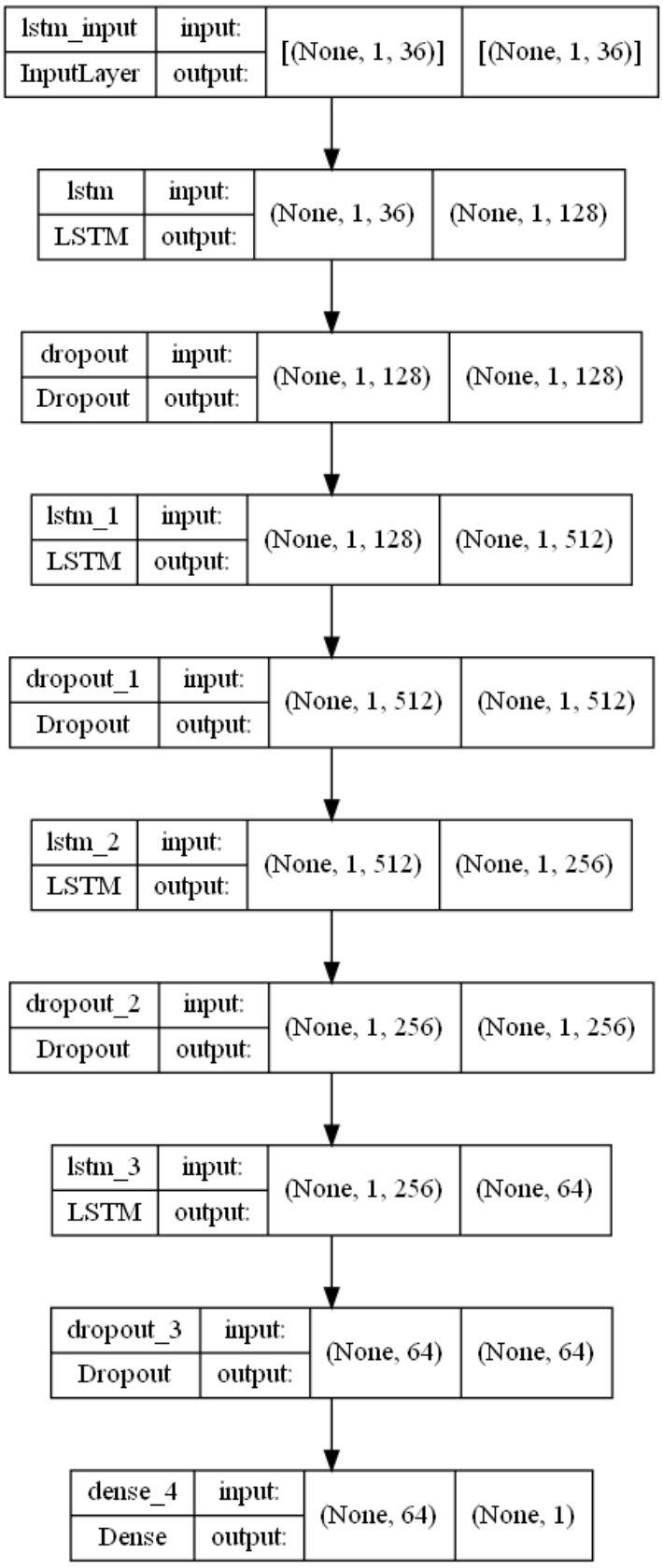

Figure 6: LSTM Model Architecture

6 Results

After all of the experiments were run, each algorithm was evaluated using the following metrics R-Squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). While it is argued what metric is the best for evaluating a model's performance these four were ultimately used [53]. R^2 is a useful metric that is defined as, "variance in the dependent variable that is predictable from the independent variables" [54]. In other words, this value shows the correlation between the input and output variables with the correlation being higher the closer to 1 the value is. The rest are rather straightforward, MAE is the mean difference between the actual and predicted values, MSE is the mean squared difference between the two variables, and RMSE is an MSE that has had the square root taken. The values for each algorithm are shown in Table 4.

Algorithm	**R^2**	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared Error**
Random Forest	0.8332	0.0340	0.0023	0.0479
Extreme Gradient Boosting	0.7288	0.0454	0.0037	0.0611
Decision Trees	0.6362	0.0472	0.0050	0.071
Long Short Term Memory	0.5803	0.0725	0.0090	0.0947
Gradient Boosting	0.4056	0.0705	0.0082	0.0904
Multilayer Perceptron	0.3342	0.0939	0.0142	0.1192
Linear Regression	0.1264	0.0877	0.0120	0.1096
Feed Forward Network	-0.0038	0.0937	0.0138	0.1175
Support Vector Machine	-0.0646	0.0965	0.0147	0.1210
K Means	-1067	3.1696	14.6986	3.8339
K Nearest Neighbors	NULL	NULL	NULL	NULL

Table 4: Accuracy Metrics of all Algorithms Tested

The results of this study were very interesting. Looking at the results it is clear that random forest placed at the top of all of the algorithms tested. It was expected that random forest would perform well given that it was one of if not the most commonly used algorithm for the topic of land degradation prediction. A few key observations are that both the LSTM and the MLP models performed decently but fell far short of the threshold set by random forest. While deep learning is a powerful tool and one which works well in several cases, the top two machine learning models used were far better for this type of problem.

While it is possible that deep learning could have further use for this type of problem it did not even closely match random forest in this study.

Looking closely at the top algorithms another interesting observation is their performance overall. Random forest placed highest and decision trees placed third by R^2 values. This is not too surprising seeing as how random forest is an ensemble method that uses decision trees itself and with its current use for this type of problem it should be expected that both would place at the top. The interesting part is how gradient boosting and extreme gradient boosting performed so well on the data with even the normal default gradient boosting outperforming the MLP model. Gradient boosting algorithms similar to random forest also employ decision trees to make their predictions. Therefore, the top three algorithms all used different forms of decision tree implementations to achieve their results.

Based on the performance of the models it is clear when predicting variables related to land degradation, models that make use of decision trees tend to outperform other methods. This is not an unusual occurrence. A similar paper focused on image classification with data taken from GeoTIFF images, similar to those used in this study [55]. That study found that decision trees also outperformed their neural network models [55]. From this study as well as that of this thesis, it can be discerned that despite the trend towards deep learning some traditional machine learning methods still have their uses for certain problems.

Finally, given their positions as the two top-performing algorithms, One final metric for both the random forest and extreme gradient boosting models was calculated. While the previous metrics are a good indicator of accuracy this metric puts it in a more human-friendly form of a Mean Absolute Percent Error (MAPE). This measurement gives the final absolute accuracy of the model in a percentage form which is pleasing to the eye. The final accuracies rounded to two decimal points are 13.56% for extreme gradient boosting and 10.52% accuracy for random forest.

7 Conclusion

In conclusion, artificial intelligence is a method that allows us to often make great observations from proverbial mountains of data. However, not every AI algorithm be it deep or machine learning is appropriate for every type of problem, as this research has shown. Despite being newer and arguably more complex deep learning was outperformed by decision tree based models for the prediction of land degradation, in this study. This shows that while deep learning still has many uses for some types of problems the more established machine learning algorithms can still make more accurate predictions and are therefore the better choice for the prediction of land degradation.

8 Future Work

Given the results of this study, a few more questions arise that could set the groundwork for future research. Firstly, given the proximity between random forest and extreme gradient boosting it could prove worthwhile to further explore if extreme gradient boosting can be fine-tuned to outperform random forest. As stated previously, eventually the extreme gradient boosting used in this study could not be improved further due to hardware limitations; however, running algorithms on more powerful machinery could allow it to outperform random forest.

Another possible application would be to explore why decision trees seem to vastly outperform neural networks, for this type of problem. Deep learning is often talked about as the future of AI, and its performance outperforms machine learning in many other works. However, this is not the case here as demonstrated by the final results. Examining why this is the case might prove an interesting topic for research in the future.

References

[1] H. Xie, Y. Zhang, Z. Wu, and T. Lv, "A bibliometric analysis on land degradation: Current status, development, and future directions," *Land*, vol. 9, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2073-445X/9/1/28

[2] N. G. Pricope, G. Husak, D. Lopez-Carr, C. Funk, and J. Michaelsen, "The climate-population nexus in the east african horn: Emerging degradation trends in rangeland and pastoral livelihood zones," *Global Environmental Change*, vol. 23, no. 6, pp. 1525–1541, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959378013001738

[3] K. Gashu and Y. Muchie, "Rethink the interlink between land degradation and livelihood of rural communities in chilga district, northwest ethiopia," *Journal of*

Ecology and Environment, 2018. [Online]. Available: https://jecoenv.biomedcentral.com/articles/10.1186/s41610-018-0077-0#citeas

[4] G. R. Megerssa and Y. B. Bekere, "Causes, consequences and coping strategies of land degradation: evidence from ethiopia," *Journal of Degraded and Mining Lands Management*, vol. 7, no. 1, p. 1953, 2019.

[5] N. G. Pricope, K. L. Mapes, K. M. Mwenda, S. H. Sokolow, and D. Lopez-Carr, "A review of publicly available geospatial datasets and indicators in support of drought monitoring," 2021. [Online]. Available: https://www.tools4ldn.org/resources

[6] S. Yousefi, H. R. Pourghasemi, M. Avand, S. Janizadeh, S. Tavangar, and M. Santosh, "Assessment of land degradation using machine-learning techniques: A case of declining rangelands," *Land Degradation & Development*, vol. 32, no. 3, pp. 1452–1466, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ldr.3794

[7] P. Nzuza, A. Ramoelo, J. Odindi, J. M. Kahinda, and S. Madonsela, "Predicting land degradation using sentinel-2 and environmental variables in the lepellane catchment of the greater sekhukhune district, south africa," *Physics and Chemistry of the Earth, Parts A/B/C*, p. 102931, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474706520303776

[8] T.-G. Vågen, L. A. Winowiecki, A. Abegaz, and K. M. Hadgu, "Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in ethiopia," *Remote Sensing of Environment*, vol. 134, pp. 266–275, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425713000850

[9] A. Torabi Haghighi, H. Darabi, Z. Karimidastenaei, A. A. Davudirad, S. Rouzbeh, O. Rahmati, F. Sajedi-Hosseini, and B. Klöve, "Land degradation risk mapping using topographic, human-induced, and geo-environmental variables and machine learning

algorithms, for the pole-doab watershed, iran," *Environmental Earth Sciences*, vol. 80, no. 1, p. 1, Jan 2021. [Online]. Available: https://doi.org/10.1007/s12665-020-09327-2

[10] S. Cerretelli, L. Poggio, A. Gimona, G. Yakob, S. Boke, M. Habte, M. Coull, A. Peressotti, and H. Black, "Spatial assessment of land degradation through key ecosystem services: The role of globally available data," *Science of The Total Environment*, vol. 628-629, pp. 539–555, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969718304741

[11] D. I. Rukhovich, P. V. Koroleva, D. D. Rukhovich, and N. V. Kalinina, "The use of deep machine learning for the automated selection of remote sensing data for the determination of areas of arable land degradation processes distribution," *Remote Sensing*, vol. 13, no. 1, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/1/155

[12] M. Stocking, "Land degradation," in *International Encyclopedia of the Social Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 8242–8247. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B008043076704184X

[13] S. A. Shahid, M. Zaman, and L. Heng, *Soil Salinity: Historical Perspectives and a World Overview of the Problem.* Cham: Springer International Publishing, 2018, pp. 43–53. [Online]. Available: https://doi.org/10.1007/978-3-319-96190-3_2

[14] J. B. Condliffe, "An african survey 1," *South African Journal of Economics*, vol. 7, no. 3, pp. 295–304, 1939. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1813-6982.1939.tb02212.x

[15] N. El-Hage Scialabba, "Chapter 14 - full-cost accounting for decision-making related to livestock systems," in *Managing Health Livestock Production and Consumption*,

N. El-Hage Scialabba, Ed. Academic Press, 2022, pp. 223–244. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012823019000026X

[16] T. Berdimbetov, Z.-G. Ma, S. Shelton, S. Ilyas, and S. Nietullaeva, "Identifying land degradation and its driving factors in the aral sea basin from 1982 to 2015," *Frontiers in Earth Science*, vol. 9, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/feart.2021.690000

[17] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.

[18] P. M. Chanal, M. S. Kakkasageri, and S. K. S. Manvi, "Chapter 7 - security and privacy in the internet of things: computational intelligent techniques-based approaches," in *Recent Trends in Computational Intelligence Enabled Research*, S. Bhattacharyya, P. Dutta, D. Samanta, A. Mukherjee, and I. Pan, Eds. Academic Press, 2021, pp. 111–127. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128228449000098

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

[20] A. Ali, M. Hamraz, P. Kumam, D. M. Khan, U. Khalil, M. Sulaiman, and Z. Khan, "A k-nearest neighbours based ensemble via optimal model selection for regression," *IEEE Access*, vol. 8, pp. 132 095–132 105, 2020.

[21] A. Haara and A. Kangas, "Comparing k nearest neighbours methods and linear regression–is there reason to select one over the other?" *Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS)*, vol. 4, no. 1, pp. 50–65, 2012.

[22] D. H. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," 2020.

[23] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 12 2013.

[24] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of xgboost," 11 2019.

[25] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. [Online]. Available: https://doi.org/10.1145%2F2939672.2939785

[26] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," vol. 2049, 01 2001, pp. 249–257.

[27] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 14 – 23, 01 2011.

[28] E. Gültekin, H. K. Kaynak, and H. Çelik, "Decision tree regression method application for prediction of woven fabrics tear strength," 12 2021.

[29] H. Zhang, H. Yu, Y. Li, and B. Hu, "Improved k-means algorithm based on the clustering reliability analysis," in *Proceedings of the 2015 International Symposium on Computers Informatics*. Atlantis Press, 2015/01, pp. 2516–2523. [Online]. Available: https://doi.org/10.2991/isci-15.2015.326

[30] I. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, 11 2021.

[31] M. Sazli, "A brief review of feed-forward neural networks," *Communications, Faculty Of Science, University of Ankara*, vol. 50, pp. 11–17, 01 2006.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[33] M.-C. Popescu, V. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, 07 2009.

[34] S. Pal, R. Chakrabortty, A. Arabameri, M. Santosh, A. Saha, I. Chowdhuri, P. Roy, and M. Shit, "Chemical weathering and gully erosion causing land degradation in a complex river basin of eastern india: an integrated field, analytical and artificial intelligence approach," *Natural Hazards*, vol. 110, 01 2022.

[35] A. Yacine, Z. Saidani, R. Touati, Q. Pham, S. Pal, M. Firuza, and F. Balik Sanli, "Assessing landslide susceptibility using a machine learning-based approach to achieving land degradation neutrality," *Environmental Earth Sciences*, vol. 80, 09 2021.

[36] V. Habibi, H. Ahmadi, M. Jaffari, and A. Moeini, "Prediction of land degradation by machine learning methods," *Earth Sciences Research Journal*, vol. 25, pp. 353–362, 10 2021.

[37] R. Chakrabortty, S. Pal, M. Sahana, A. Mondal, J. Dou, B. Pham, and A. P. Yunus, "Soil erosion potential hotspot zone identification using machine learning and statistical approaches in eastern india," *Natural Hazards*, vol. 104, 11 2020.

[38] A. Garg, I. Wani, and V. Kushvaha, "Application of artificial intelligence for predicting erosion of biochar amended soils," *Sustainability*, vol. 14, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/2/684

[39] A. Abolhasani, G. Zehtabian, H. Khosravi, O. Rahmati, E. Heydari Alamdarloo, and P. D'Odorico, "A new conceptual framework for spatial predictive modeling of land degradation in a semi-arid area," *Land Degradation & Development*, vol. n/a, no. n/a. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ldr.4391

[40] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017, big Remotely Sensed Data: tools, applications and experiences. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425717302900

[41] M. Blumenthal, E. Grover-Kopec, M. Bell, and J. del Corral, "The iri/ldeo climate data library: Helping people use climate data," *AGU Fall Meeting Abstracts*, 12 2005.

[42] C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen, "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes," *Scientific Data*, 2015. [Online]. Available: https://www.nature.com/articles/sdata201566#citeas

[43] C. Funk, P. Peterson, S. Peterson, S. Shukla, F. Davenport, J. Michaelsen, K. R. Knapp, M. Landsfeld, G. Husak, L. Harrison, J. Rowland, M. Budde, A. Meiburg, T. Dinku, D. Pedreros, and N. Mata, "A high-resolution 1983–2016 tmax climate data record based on infrared temperatures and stations by the climate hazard center," *Journal of Climate*, vol. 32, no. 17, pp. 5639 – 5658, 01 Sep. 2019. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/32/17/jcli-d-18-0698.1.xml

[44] Didan, "Modis/terra vegetation indices 16-day l3 global 500m sin grid v061," 2021. [Online]. Available: https://doi.org/10.5067/MODIS/MOD13A1.061

[45] A. Huete, K. Didan, T. Miura, E. Rodriguez, X. Gao, and L. Ferreira, "Overview of the radiometric and biophysical performance of the modis vegetation indices," *Remote Sensing of Environment*, vol. 83, no. 1, pp. 195–213, 2002, the Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface

Monitoring. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425702000962

[46] V. Urrea, A. Ochoa, and O. Mesa, "Seasonality of rainfall in colombia," *Water Resources Research*, vol. 55, no. 5, pp. 4149–4162, 2019. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023316

[47] A. Esquivel, L. Llanos-Herrera, D. Agudelo, S. D. Prager, K. Fernandes, A. Rojas, J. J. Valencia, and J. Ramirez-Villegas, "Predictability of seasonal precipitation across major crop growing areas in colombia," *Climate Services*, vol. 12, pp. 36–47, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405880718300177

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[50] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[51] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[52] I. L. Alsammak, H. M. A. Sahib, and W. H.Itwee, "An enhanced performance of k-nearest neighbor (k-NN) classifier to meet new big data necessities," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, p. 032013, nov 2020. [Online]. Available: https://doi.org/10.1088/1757-899x/928/3/032013

[53] T. Chai and R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?– arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, pp. 1247–1250, 06 2014.

[54] D. Chicco, M. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 07 2021.

[55] J. Treboux, D. Genoud, and R. Ingold, "Decision tree ensemble vs. n.n. deep learning: Efficiency comparison for a small image dataset," in *2018 International Workshop on Big Data and Information Security (IWBIS)*, 2018, pp. 25–30.

APPENDIX

BIOGRAPHICAL SKETCH