BAMBOO CLASSIFICATION USING BOOSTING ENSEMBLES AND VOTING ALGORITHMS

Mini Project I report submitted in partial fulfillment of the Requirements for the Award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Uday Kiran Mogalapu	208W1A05A0
Yaddanapudi Naga Himaja	208W1A05C8

Under the Guidance of

Dr. K. SRINIVAS, M.Tech, Ph.D

Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
V.R SIDDHARTHA ENGINEERING COLLEGE
Autonomous and Approved by AICTE, NAAC A+, NBA Accredited
Affiliated to Jawaharlal Nehru Technological University, Kakinada
Vijayawada 520007

April, 2023

VELAGAPUDI RAMAKRISHNA SIDDHARTHA ENGINEERING COLLEGE

(Autonomous, Accredited with 'A+' grade by NAAC)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the Mini project I course report entitled **" BAMBOO CLASSIFICATION USING BOOSTING ENSEMBLES AND VOTING ALGORITHMS"** being submitted by

Uday Kiran Mogalapu	208W1A05A0
Yaddanapudi NagaHimaja	208W1A05C8

in partial fulfilment for the award of the Degree of Bachelor of Technology in Computer Science and Engineering to the Jawaharlal Nehru Technological University, Kakinada, is a record of bonafide work carried out during the period from 2023 - 2024.

Dr. K. Srininvas M.Tech, Ph.D	**Dr.D.Rajeswara Rao,** M.Tech, Ph.D
Professor & Guide	**Professor & HOD,CSE**

DECLARATION

We hereby declare that the Mini Project I project entitled "BAMBOO CLASSIFICATION AND MAPPING USING SUPERVISED LEARNING TECHNIQUES" submitted for the B.Tech Degree is our original work and the dissertation has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles.

Place: Vijayawada — Uday Kiran Mogalapu (208W1A05A0)

Date: — Yaddanapudi Naga Himaja (208W1A05C8)

ACKNOWLEDGEMENT

We would like to thank **Dr. A. V. Ratna Prasad**, Principal of Velagapudi Ramakrishna Siddhartha Engineering College for the facilities provided during the course of Mini Project.

We have been bestowed with the privilege of thanking **Dr. D. Rajeswara Rao**, Professor and Head of the Department for his moral and material support.

We would like to express our deep gratitude to our guide **Dr. K. Srinivas**, Professor for his persisting encouragement, everlasting patience and keen interest in discussion and for his numerous suggestions which we had at every phase of this project.

We owe our acknowledgements to an equally long list of people who helped us in mini project 1 work.

Place: Vijayawada	Uday Kiran Mogalapu (208W1A05A0)
Date:	Yaddanapudi NagaHimaja (208W1A05C8)

Abstract

Boosting ensembles are commonly used to enhance the accuracy of classification models by training multiple models sequentially and merging their predictions. However, selecting the best model from the ensemble can be challenging. Voting algorithms are a popular method for selecting the best-fit model by aggregating the predictions of multiple models. However, existing voting algorithms have limitations, including hard voting being sensitive to class imbalance and soft voting assuming that all models are equally reliable. Weighted voting is also challenging, requiring expert knowledge and experience to assign weights to models. In this Mini project-1, firstly four boosting ensembles are trained using SVM, CART, Decision Trees, and Naive Bayes algorithms as base classifiers. Later, apply voting algorithms such as hard voting, soft voting, weighted voting, and others on these four ensembles. Accuracy is taken as the metric for evaluating the performance. Dataset is collected from two bamboo-dominated districts named Dima Hasao and Karbi Anglong, one as a training site and the other as a testing site in Assam state.

Keywords: Boosting ensembles, Voting algorithms, Best-fit model, Supervised learning algorithms, Bamboo classification, Accuracy

Table of Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Basic Concepts

Image Classification:

The process of classifying and then labelling groups of pixels or vectors within an image in accordance with predetermined guidelines is known as image classification. Generally, it is divided into two groups: supervised and unsupervised image categorization methods.
Stages of Image Classification are:

- **Data Acquisition:** Data must initially be obtained from the appropriate sources before it can be saved, cleaned up, put through preprocessing, or used for other activities.
- **Image Pre-processing:** This technique seeks to improve the picture data (features) by decreasing unwanted distortions and improving a few key image quality.

Supervised Learning Techniques:

- **CART (Classification And Regression Trees):** The supervised method known as CART (Classification and Regression Trees) is used to solve classification and regression problems. What the aim variable is matters.When utilising classification, we try to forecast a class label; as a result, the output is limited to a narrow range of potential values, such as whether or not it will rain tomorrow. Any number of values, such as the cost of a piece of property, can be included in the output when regression is used to forecast a numerical label.The threshold value of each characteristic determines how the decision tree's nodes are broken into subnodes.The highest homogeneity subnodes are found using the CART method using the Gini Index criterion.
- **SVM (Support Vector Machine):** The fundamental idea behind this approach is to identify the class border that minimises the travel time to the closest data point. The maximum-margin hyperplane, which serves as a border, is computed using support vectors created from these locations. The dimensionality of this hyperplane is determined by the amount of features

to be used for categorization. The decision surface, which accentuates the boundary between classes, is the main focus of SVM sections. The data points close to these surfaces are referred to as support vectors, and they are known as ideal hyperplanes.
figure 1.1[1] shows the implementation of SVM

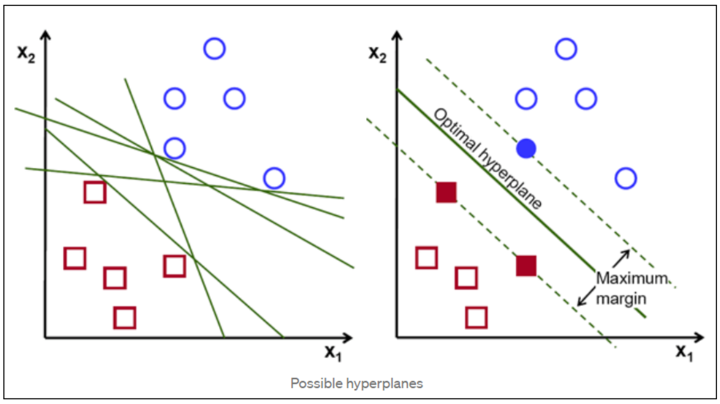

Figure 1.1: Support Vector Machine

- **Naive Bayes:** The supervised learning algorithm known as Naive Bayes, which addresses classification problems, is made out of the words Naive and Bayes. It is foolish to assume that the presence of one feature has no bearing on the incidence of other traits. It depends on the Bayes' Theorem premise, says Bayes. It bases its predictions on the possibility that an object truly exists because it is a probabilistic classifier.
  figure 1.2[2] shows the Bayesian Formula

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Figure 1.2: Bayesian Formula

- **Random Forest:** One of the supervised machine learning methods, the Random Forest algorithm, is commonly used to handle classification and regression problems. We are aware that a forest is made up of numerous different species of trees, and that the more diversified the tree species, the more robust the forest will be. Similar to this, the Random Forest Algorithm improves in accuracy and problem-solving capacity as the number of trees grows. More decision trees are applied by Random Forest to various input data subsets to increase the dataset's prediction accuracy.
  One of the most important characteristics of the Random Forest Algorithm is its capacity to handle data sets containing both continuous variables, as in

regression, and categorical variables, as in classification. It produces better results when it comes to classification problems. In order to increase the estimated accuracy of the input dataset, the Random Forest classifier combines the results from several decision trees applied to different subsets of the input dataset. Overfitting should be avoided as accuracy increases because there are more trees in the forest.

figure 1.3[3] shows the implementation of RF

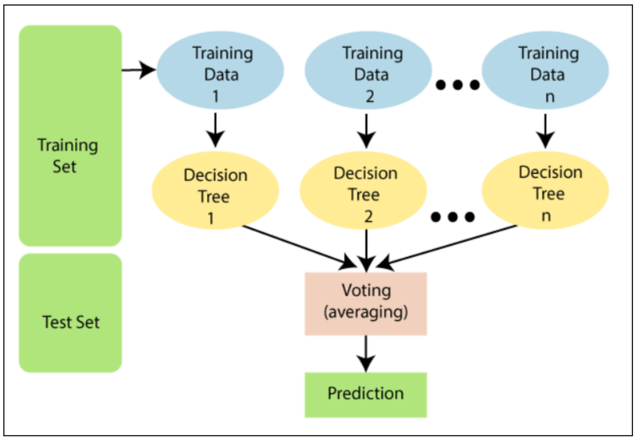

Figure 1.3: Random Forest

Boosting:

Models are trained successively in boosting, with each model being taught to fix the flaws of the one before it.By using weighted voting, where each model's contribution is decided by how well it performs on the training set, the final prediction is obtained.Boosting is an iterative procedure that modifies the training instance weights in response to faults in the prior model.0

figure 1.4[4]shows the Boosting Technique

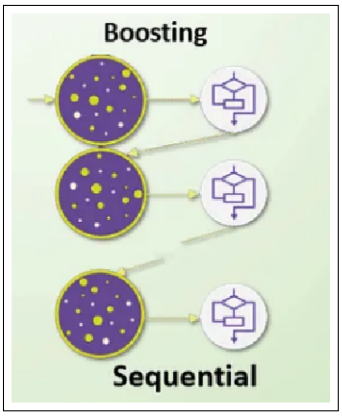

Figure 1.4: Boosting Technique

Voting Algorithm:

In machine learning, a voting classifier is a particular form of ensemble approach that combines the predictions of various models (classifiers) to get a final prediction. In this instance, a voting method is used to integrate the AdaBoost ensemble classifiers for individual classifiers into a single classifier. figure 1.5[5] shows how Voting algorithm works

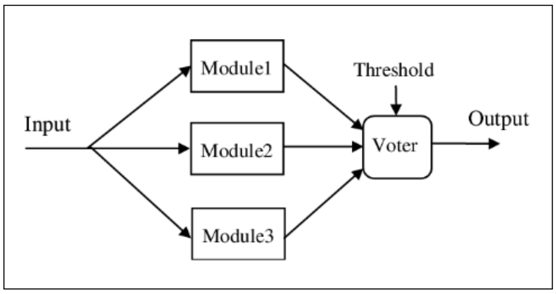

Figure 1.5: Voting Algorithm

1.2 Problem Statement

To explore the effectiveness of voting algorithms in selecting the best-fit model from four boosting ensembles trained with SVM, CART, Decision Trees, and Naive Bayes algorithms as base classifiers. The project aims to evaluate different voting algorithms, including hard voting, soft voting, weighted voting, and others, based on their impact on classification accuracy. The dataset used for this project is collected from two bamboo-dominated districts, Dima Hasao and Karbi Anglong, in Assam state. The project will utilize satellite images obtained from Sentinel-2, a mission under the Copernicus Programme, which offers high-resolution imagery ranging from 10 meters to 60 meters. By leveraging the capabilities of remote sensing technology, the project seeks to accurately classify the region.

1.3 Objectives

The Objective of our project is:

- To train four boosting ensembles by using SVM, CART, Decision Trees, and Naive Bayes algorithms as base classifiers.
- To apply voting algorithms such as hard voting, soft voting, weighted voting, and others on these four boosting ensembles.
- To Classify bamboo as Pure, Dominated, or Mixed, using accuracy as the primary performance metric.

1.4 Scope

The scope of our projecct is

- Limited to Karbi Anglong East West district and Dima Hasao District of North-Eastern Region of India as study area.
- Confined to four supervised learning techniques and boosting ensemble technique.

1.5 Advantages

- Automated classification processes save time and effort compared to manual classification, making data analysis and decision-making more efficient.
- Classification models can uncover hidden patterns and relationships within the data, providing valuable insights for further analysis and understanding.
- Once trained, classification models can be used to make predictions on new, unseen data, facilitating forecasting and proactive decision-making.

1.6 Applications

Various applications are present for classification of bamboo. Some of them are as follows

- Environmental impact assessment
- Illegal activity detection
- Bamboo resource management
- Forest inventory assessment

Chapter 2

LITERATURE REVIEW

This chapter contains the list of journals that we have studied under literature survey. We focused on the approaches for maintaining accuracy in these papers. Our motive was to find the best research that would help in building our project.

2.1 Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students

In order to create a model for predicting student success, D.L.N.Hoang et al. presented [6] a system in 2022 that involved gathering educational data from a learning management system. The study makes use of classifiers like Decision Trees, Naive Bayes, K-Nearest Neighbour, and Support Vector Machines. To increase the precision of the student performance model, ensemble approaches like bagging, boosting, and voting were applied. The model has behavioural characteristics and can help with decision-making to enhance the educational system. The study shows the validity of the suggested model in predicting student accomplishment with an accuracy of 96.8.

Advantages:

Voting is done along with ensemble approaches like Bagging and Boosting to increase the model's accuracy.It utilises a sizable dataset from an LMS, which improves the precision and dependability of the outcomes.

Disadvantages:

Relies on labeled data, and it may not be suitable for all types of data. Uses data only from a single learning management system, which may limit the generalizability.

2.2 Scene Classification of Remotely Sensed Images using Ensembled Machine Learning Model

In 2021, Deepan Perumal et al. proposed [7] a model that includes an ensemble technique to improve the precision and dependability of remote sensing picture scene classification. The strategy entails developing several models, then merging predictions from each model to enhance performance. The paper employs many conventional classifiers in addition to a feature extraction method based on speed-up robust feature approaches to accomplish this goal. Five distinct ensemble approaches are used to integrate the conclusions of the three classifiers that perform the best. The majority voting ensemble technique using MLP, SVM-linear, SVM-kernel, and RF classifiers obtains the maximum accuracy of 93.5 when the suggested ensemble models are tested on 8000 remote sensing images. The study finds that ensembling methods can considerably increase the accuracy and dependability of remote sensing picture scenes, and it advises applying deep learning methods to make even more advancements.

Advantages:

Applies feature extraction technique using speed-up robust feature techniques Proposes an ensemble approach for remote sensing image classification.

Disadvantages:

Uses only traditional classifiers for the base models and does not explore the use of deep learning models. Does not provide a detailed analysis of the computational requirements.

2.3 Voting-Based Ensemble Learning Algorithm for Fault Detection in Photovoltaic Systems under Different Weather Conditions

In 2022, N. C. Yang et al. proposed [8] a EL-VLR-DT-SVM model which uses a voting-based ensemble learning algorithm incorporating linear regression, decision tree, and support vector machine for detecting and diagnosing photovoltaic (PV) faults. Accuracy, precision, recall, F1 score, and computational time are evaluated as performance metrics. On all metrics, the suggested approach performs better than other algorithms, with accuracy values in Scenario 1-3 of 99.89, 100, and 100, respectively. The suggested method performs well in identifying and diagnosing PV faults.

Advantages:

Uses a feature selection method to identify the most important features for fault detection.The dataset used in the study is publicly available making the proposed approach more robust for real-world applications.

Disadvantages:

The accuracy of the proposed approach might vary depending on the location of the PV system and the climate conditions in that area.Does not provide a clear explanation of how the model makes predictions.

2.4 Coastal Wetland Mapping Using Ensemble Learning Algorithms: A Comparative Study of Bagging, Boosting and Stacking Techniques

L. Wen et al. proposed [9] in 2020. The wetland/non-wetland pixel labels utilised in this work were collected from sentinel-2 images and field observations. The authors employed several combinations of characteristics extracted from remote sensing data to train and evaluate the models. They examined the three most popular EL methods for mapping wetland distribution: bagging, boosting, and stacking. With an overall accuracy of 97.8

Advantages:

It emphasises the possibilities of adopting alternative ensemble learning techniques in remote sensing applications, such as XGBoost and bagged trees.By selecting only highly credible wetland polygons for sampling, the study ensured the accuracy of its results.

Disadvantages:

Sentinel-2 Fractional Cover data are used in the study for wetland mapping, although only two of the three components are used, which reduces the accuracy.According to the study, ensemble approaches had limited utility for mapping unusual or uncommon wetland types since they were less effective at differentiating minority classes.

2.5 Satellite Image Classification Using a Hierarchical Ensemble Learning and Correlation Coefficient-Based Gravitational Search Algorithm

K. Thiagarajan et al. suggested [10] a new technique for classifying satellite images in 2021 by combining a hierarchical framework, ensemble learning, and the best possible feature selection. The CCGSA Algorithm is used to extract the best features from the feature set while the HFEL uses three different forms of CNN to extract features from images in a hierarchical fashion. The suggested approach is examined for accuracy, precision, and recall using data from the SAT-4, SAT-6, and Eurosat datasets. The findings demonstrate that the suggested HFEL-CCGSA approach works better than conventional CNNs like AlexNet, LeNet-5, and ResNet, obtaining a classification accuracy of 99.99

Advantages:

The accuracy of classification is increased by combining multiple CNNs (AlexNet, LeNet-5, and ResNet) with the best feature selection algorithm (CCGSA).The exact classification of satellite pictures is made possible by the union of HFEL and CCGSA.

Disadvantages:

The employment of several CNNs and feature selection methods in the suggested method could result in a considerable computational resource requirement.The proposed method is not contrasted with cutting-edge techniques on datasets other than SAT-4, SAT-6, and Eurosat in the study.

2.6 Performance of Machine Learning-Based Multi-Model Voting Ensemble Methods for Network Threat Detection in Agriculture 4.0

N. Peppes et al. [11] proposal from 2021 emphasises the use of network traffic classification to detect and stop harmful network activities. The performance of five distinct machine learning (ML) classifiers—K-Nearest Neighbours (KNN), Support Vector Classification (SVC), Decision Tree (DT), Random Forest (RF), and Stochastic Gradient Descent (SGD)—is evaluated in the study. A hard voting and a soft voting ensemble model comprising these classifiers are also suggested. Three versions of the NSL-KDD dataset were used for the evaluation: the baseline dataset, the undersampled dataset, and the oversampled dataset. The outcomes demonstrate that the ensemble models outperform the individual models.

Advantages:

The authors have used a multi-model voting ensemble method that combines the results of several machine learning modelsThe paper uses real-world data collected from an experimental farm, which enhances the generalizability of the results to other agricultural settings.

Disadvantages:

The paper does not provide much detail on the feature selection process and the specific features used for training the machine learning models.It does not provide any information on the computational requirements and scalability of the proposed approach.

2.7 Evaluating the effect of voting methods on ensemble-based classification

In 2020, F. Leon et al. proposed [12], which examines the impact of various voting methods on the effectiveness of two classification algorithms—k-Nearest Neighbour and Naive Bayes—and makes use of datasets with various degrees of difficulty to evaluate the effectiveness of various voting techniques in ensemble-based classification. The paper aggregates the outputs of the classifiers using four distinct voting procedures, including the plurality method, Borda count, Copeland's approach, and Single Transferable Vote (STV), and compares their performance in terms of classification accuracy. The outcomes demonstrate that the STV can be a useful replacement for the plurality approach.

Advantages:

Gives a thorough investigation of the effects of various voting techniques when applied to datasets with various degrees of complexity.The less popular voting algorithms that are really more useful for multiclass issues are explored in this study.

Disadvantages:

The computation of the preference ranking, which takes longer than a simple vote count, is the key challenge in employing the STV approach.The effect of hyperparameter adjustment on the performance of the voting techniques, which may have an impact on their overall accuracy, is not covered in the paper.

2.8 A Study on Ensemble Methods for Classification

In 2020, R. Harine Rajashree et al. proposed [13] in which they used a number of publicly available datasets for their experiments, including the Breast Cancer Wisconsin (Diagnostic) dataset, the Ionosphere dataset, the Mushroom dataset, and the German Credit dataset, among others. Description: This study focuses on ensemble learning. Bagging, boosting, and stacking are the three primary categories of ensemble approaches, each of which has a number of underlying algorithms. Machine learning issues including idea drift, bias variance error, and class imbalance are reduced using ensembles.Although ensembles have shown to increase accuracy, their use can be costly in terms of storage and time, and it is yet unclear how factors like ensemble size and learner selection will affect accuracy.This study intends to examine these methods' applications, advancements, and classification-related difficulties.

Advantages:

Analysis on different ensemble methods on classification accuracy and the impact of dataset characteristics on ensemble performance.The paper provides a clear explanation of the ensemble methods used, including bagging, boosting, and stacking, as well as the base classifiers.

Disadvantages:

The paper lacks discussion on the computational efficiency and complexity of the ensemble methods used.Study is limited to a specific set of machine learning algorithms and datasets, which may not be representative of all possible scenarios.

Chapter 3

ANALYSIS AND DESIGN

This chapter includes the analysis of requirements for the proposed project. This chapter contains

- Functional Requirements.
- Non-Functional Requirements.

3.1 Functional Requirements

In order to make sure that the hardware and software requirements meet the necessary and applicable standards, conducting a functional requirement analysis entails thoroughly going over, assessing, and documenting the needs. Functional requirements go through numerous stages. analysis.
The Functional Requirements include:

Software Requirements

- Google Earth Engine(GEE) Code Editor:
  An Earth Engine online IDE The Earth Engine (EE) Code Editor is the name given to the JavaScript API.The code editor's features enable the quick and simple development of intricate geospatial operations. The Code Editor has features like a JavaScript code editor, a GIT-based script manager, an output console, a task manager to handle lengthy queries, an interactive map query, a search of the saved scripts or the data archive, and geometry drawing tools. It also has features that help visualise the images of geospatial data.
- Copernicus - Sentinel 2 data:
  The Copernicus Open Access Hub (formerly known as Sentinels Scientific Data Hub) offers comprehensive, open, and free access to Sentinel-1, Sentinel-2, Sentinel-3, and Sentinel-5P user products starting with the In-Orbit Commissioning Review.
- Java Script (Programming language):
  When classifying photos, especially satellite photographs, the Google Earth code editor is frequently utilised. Java script is the programming

language employed in this editor. When using satellite imagery, Java script is a very simple programming language. The java script built-in methods that we utilised to accomplish categorization are as follows:

  – smileNaiveBayes
  – libsvm
  – smileCart
  – smileRandomForest

- Python (Programming language):
  Python is extensively used in writing code for classification machine learning projects. Its rich ecosystem of libraries, such as scikit-learn, allows for efficient data preprocessing, model training, evaluation, and visualization. Python's easy-to-understand syntax facilitates the implementation of classification algorithms and enables seamless integration with various techniques, including feature selection, cross-validation, and hyperparameter tuning, ultimately leading to accurate and robust classification models.

Hardware Requirements

- Mac OS X 10.11, Windows 7 or 10, or another contemporary operating system
- RAM/Main Memory - 4GB DDR4 3200Mhz
- Internet Connectivity
- Configuration - Intel Core i5 and above
- x86 64 - bit CPU

3.2 Non-Functional Requirements

Non-functional requirements define the limitations necessary to carry out functionality and explain how a system should function. The quality attributes of the system are another name for these requirements.
The project's non-functional needs are as follows::

Usability:

The end user's satisfaction and acceptance are one of the key elements that determine a project's success. How easy or difficult it will be for users to comprehend and utilise a system depends on its usability. It is assessed using usefulness, intuitiveness, and low perceived workload. The user experience requirements should be considered from the start of a project.

Reliability:

Reliability measures how likely it is that a system will perform as intended for a specific amount of time. Reduced reliability is caused by bugs in the code, hardware problems, or problems brought on by other system components. You can also monitor the typical length of time the system lasts before failing or track the proportion of actions that are successfully carried out to evaluate the dependability of software and The system suitably complies with performance standards.

Performance:

Performance is a term used to describe how responsive a system responds to different user inputs. It is a mark of quality. A poor user experience is the outcome of poor performance. When the system is overloaded, its security is jeopardised. Particularly when the project is in the architecture phase, this is one of the most important factors to take into account..

Availability:

The availability is measured by how long it takes the system's features and services to be available for use in all operations. It's important to explain how the maintenance will be affected by the reduction. When defining the availability criteria, the team must describe the most crucial system components that must always be available. Users should be informed in advance if the system or one of its components becomes unavailable.

Maintainability:

A well-designed system should function for a very long time.As a result, maintenance that is both preventive and corrective will be regularly needed. The system's capacity to develop and improve its features and capabilities could be referred to as maintenance.

Cost Effective:

Cost is an important factor to take into account because it will influence how the system is created. A system should always be affordable to use and improve in order to avoid financial problems. The best initiatives are always those that can be constructed more cost-effectively.

Chapter 4

SOFTWARE DESIGN

This chapter consists of the design of the software Life Cycle model diagrams and their detailed explanation.Design is about choosing the architecture and solutions appropriate to the problem

4.1 Software Development Lifecycle: Agile Development Model

Agile Development involves breaking down a project into smaller tasks, and working on them in short, iterative cycles called sprints. It prioritize delivering working software frequently and adapting to changing requirements as needed. For our project it would be a good fit because the project is complex, and the requirements may evolve over time. This Approach would allow us to adjust the plans as new information becomes available, ensuring that the final product is of the highest quality.
The Figure 4.1[14] describes the Agile development model.

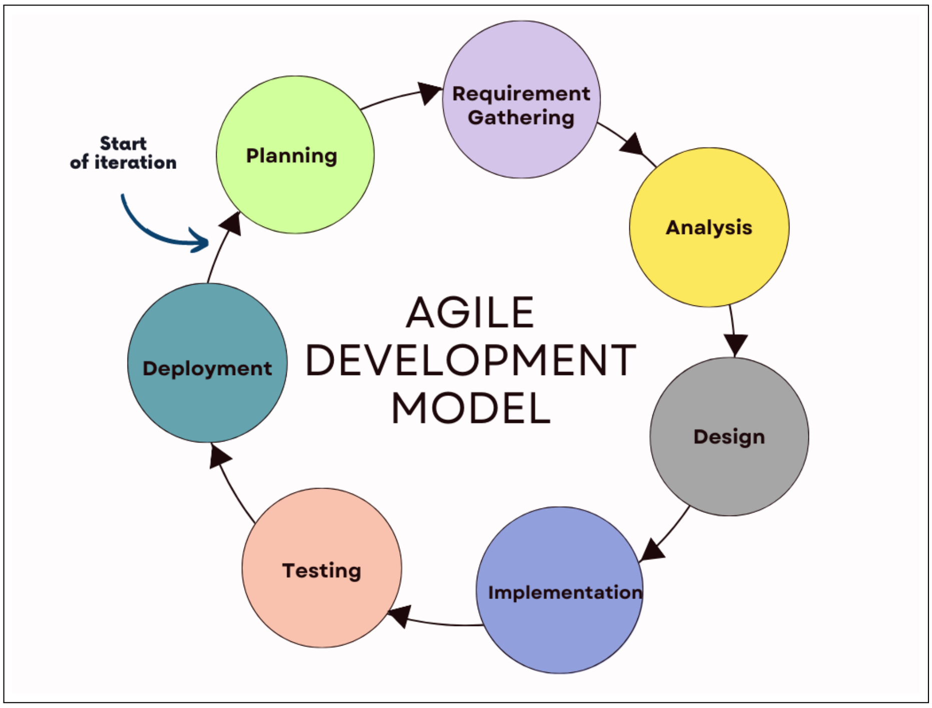

Figure 4.1: Agile model

4.2 UML Diagrams

4.2.1 Use-Case Diagram

A use case illustration illustrates the system's dynamic behaviour. It incorporates actors, use cases, and their interactions, which helps to capture the overall system's functioning. It outlines the duties, procedures, and services that a system or application subsystem requires. (Figure 4.2). In this system, Developer first collects the data and then preprocesses the data to obtain csv files from the image files. Later model is designed in a way that it takes four adaboost ensembles which are generated using four supervised learning algorithms for each ensemble. Model is trained and tested against the testing dataset. Finally the user will be displayed with the accuracy.

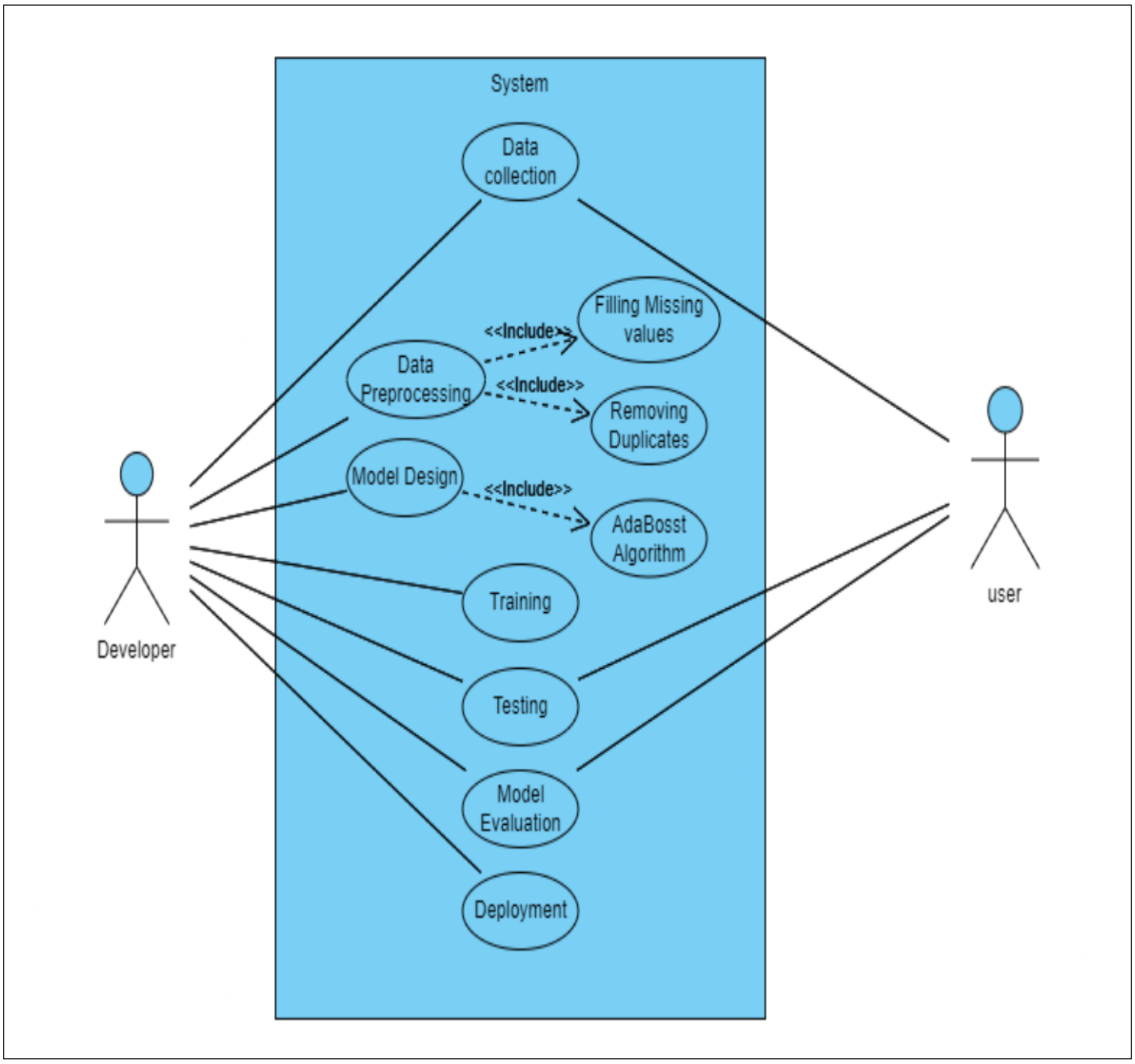

Figure 4.2: Use Case Diagram

4.2.2 Sequence Diagram

The sequence diagram, sometimes referred to as an event diagram, below illustrates the message flow through the system.It helps in visualising a variety of dynamic situations. Any interaction between two lifelines is shown as a chronologically ordered set of events, demonstrating that both lifelines were active during the communication. A dotted vertical line crosses the bottom of the page to represent the message flow, while a vertical bar symbolises the lifeline. In below system (Figure 4.3) User uploads the data to the system. Then data will be preprocessed first and then features are extracted. A model is designed using four boosting ensembles which take four supervised learning algorithms. Now model is trained and finally tested against testing dataset. Lastly, performance is evaluated and model is deployed at the used end.

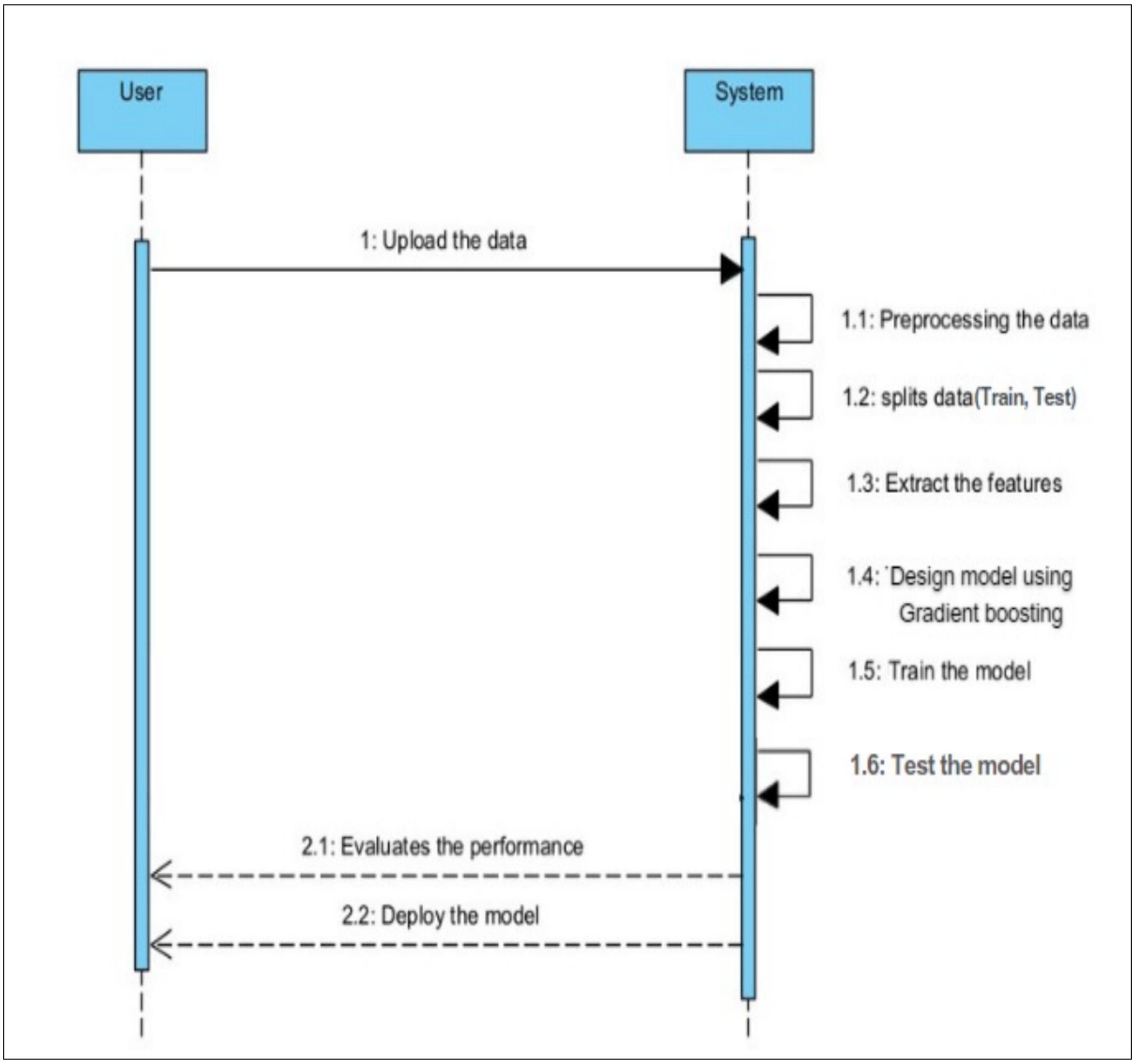

Figure 4.3: Sequence Diagram

4.2.3 Activity Diagram

A UML activity diagram outlines the actions required to complete a process in order to show how a system functions. As shown in (Figure 4.4), the flow is clear and understandable for this project. During the picture collection phase, sentinel 2 gathers satellite images. After pre-processing, the features are extracted from this data. After that, the model is trained to become more accurate using the ensemble technique of boosting. The model is afterwards examined using the testing dataset. Performance evaluation measures are then used to evaluate performance.

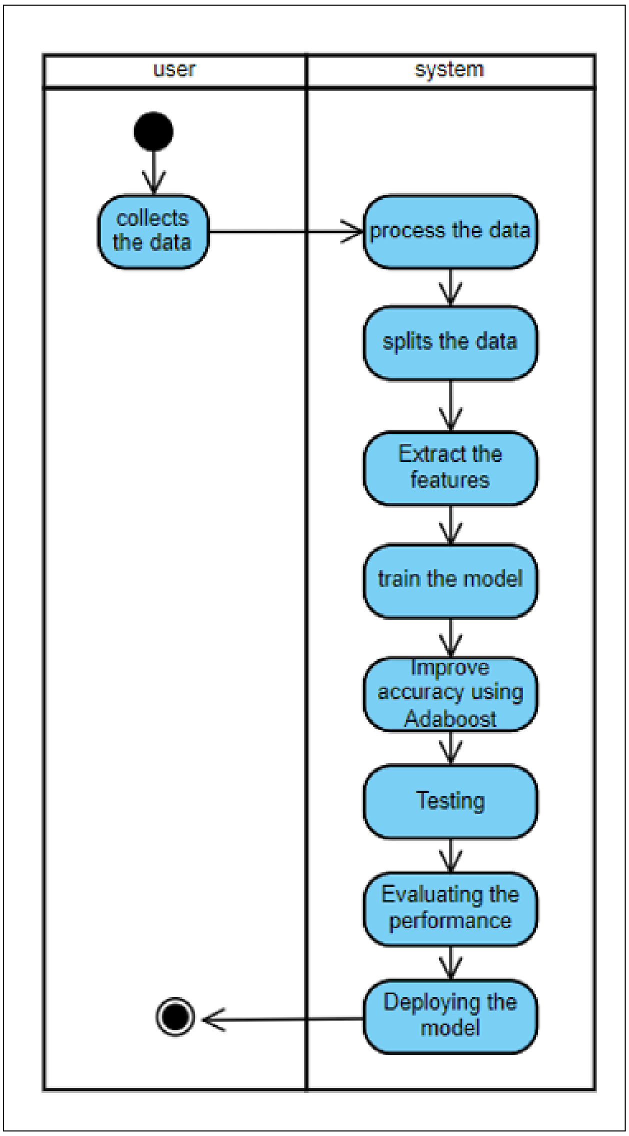

Figure 4.4: Activity Diagram

Chapter 5

PROPOSED SYSTEM

This chapter includes the proposed system architecture along with the modules of methodology and dataset collection

5.1 Architecture

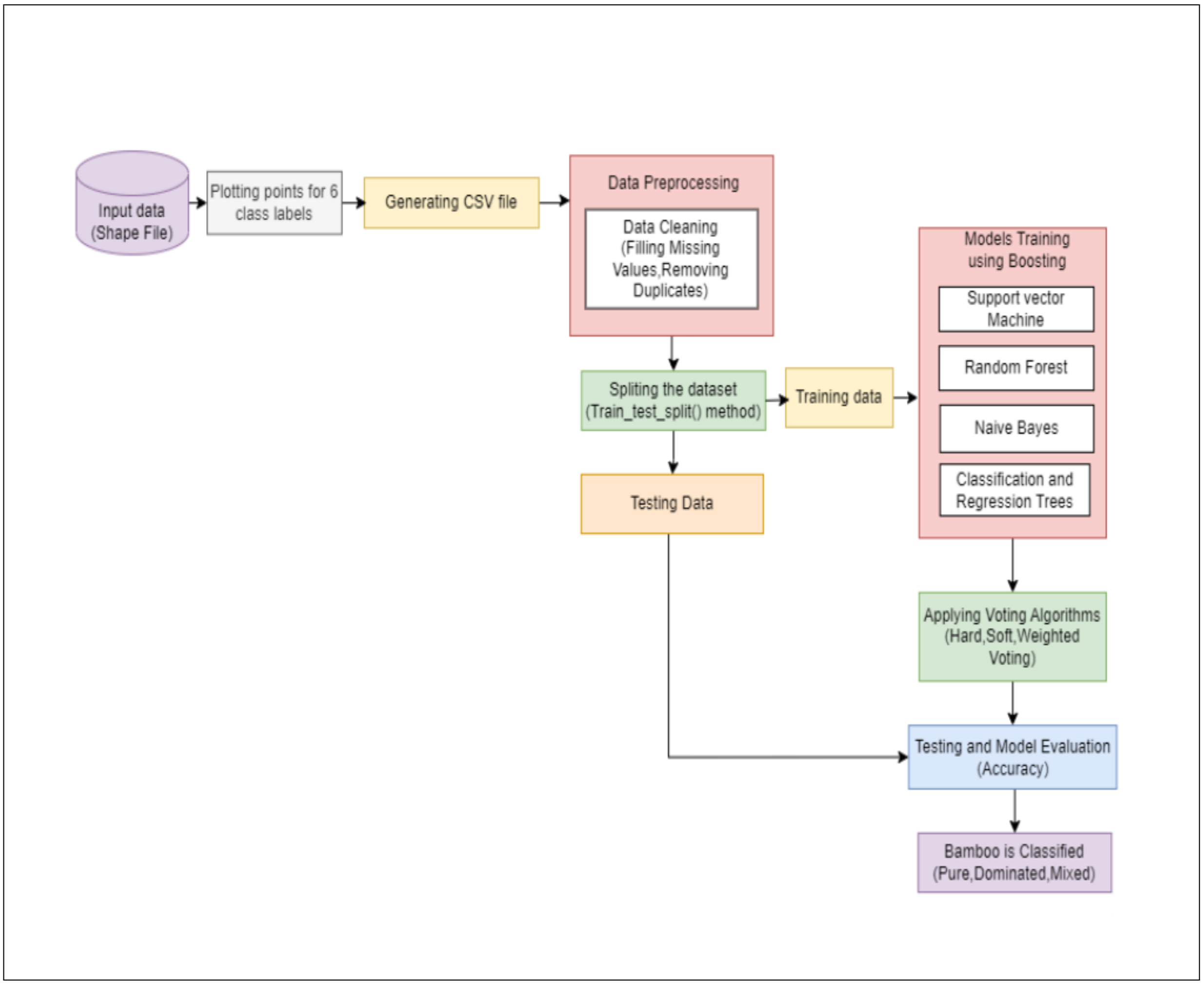

Figure 5.1: Architecture of proposed system diagram

As shown in the Architecture diagram (figure 5.1), the project is divided into four different phases.

1. Data Collection
2. Data Preprocessing
3. Model Training
4. Performance Evaluation

5.1.1 Data Collection:

Import the required district shapefile Upgrade Google Earth Code Editor map to import the district layer.Import Sentinel data from Copernicus.eu, Manually annotate the satellite imagery to create point plots, Convert the annotated data into a CSV file format

5.1.2 Data Preprocessing:

Preprocessing steps like filling missing values and removing duplicates are performed. Selecting location and label Columns and dropping other unwanted columns, Class labels considered are pure, dominated, mixed, water, forest, fallow land ,we fill the missing values by satistical measures like mean,meadian,mode and also filling them manually after the null values are filled we remove the repeated values present in the data dy using the 'drop_duplicates()' method in python

5.1.3 Model Training:

The base classifiers for enhancing ensembles are SVM, CART, Naive Bayes, and Random Forest. The AdaBoost algorithm involves the following steps:
First, we determine the sample weight and allocate it to each sample using formula as shown in figure 5.2

$$\boxed{sample\ weight = \frac{1}{\#\ of\ samples}}$$

Figure 5.2: Formula to calculate sample weight[15]

After that, we calculate the Gini Index value for each variable using the formula as shown in figure 5.3, which is done to determine which variable to use to create the first stump.

After calculating the Gini index, we calculate the amount of say using formula shown in figure 5.4, where the total error is equal to the weights of the samples

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

Figure 5.3: Formula to calculate Gini Index[16]

that were erroneously classified.

$$Amount\ of\ say = \frac{1}{2}\log\left(\frac{1 - total\ error}{total\ error}\right)$$

Figure 5.4: Formula to calculate Amount of say[17]

We now calculate the updated sample weights for the following stump using the amount of say. now Using the revised sample weights (figure 5.5) , create a boot-

$$New\ Sample\ Weight\ For\ \mathbf{Incorrect}\ Samples = Sample\ weight * e^{amount\ of\ say}$$

$$New\ Sample\ Weight\ For\ \mathbf{Correct}\ Samples = Sample\ weight * e^{-amount\ of\ say}$$

Figure 5.5: Formula to calculate New sample weight[18]

strapped dataset that includes the probabilities of each sample being selected. The procedure should be repeated n times.

Once we obtain four ensembles, we perform voting algorithms(hard, soft,weighted) to obtain final improved accuracy

- **Hard voting:**The option is chosen as the final choice if the total number of "for" votes exceeds or equals (n/2) + 1, where n is the total number of voters.
- **Soft voting:** The selection is made based on the alternative with the highest average probability.
- **Weighted voting:** The ultimate choice is made based on the alternative that received the most weighted votes.

5.1.4 Performance Evaluation:

Finally, the performance of the model is evaluated using performance evaluation metrics such as Accuracy, Recall Value. On analyzing the results, we can understand that the model with highest accuracy corresponds to the best performing model.

The accuracy can be computed using the formula:

Accuracy is equal to Number of correct predictions divided by Total number of predictions

The formula to calculate the recall (also known as sensitivity) is:

Recall = TP / (TP + FN)

F1 Score calculated as:

F1 score = 2 x [(Precision x Recall) / (Precision + Recall)].

Precision can be calculated using the formula:

Precision = TP / (TP + FP)

Figure 5.6[19]depicts the flow of AdaBoost Algorithm used in this project.

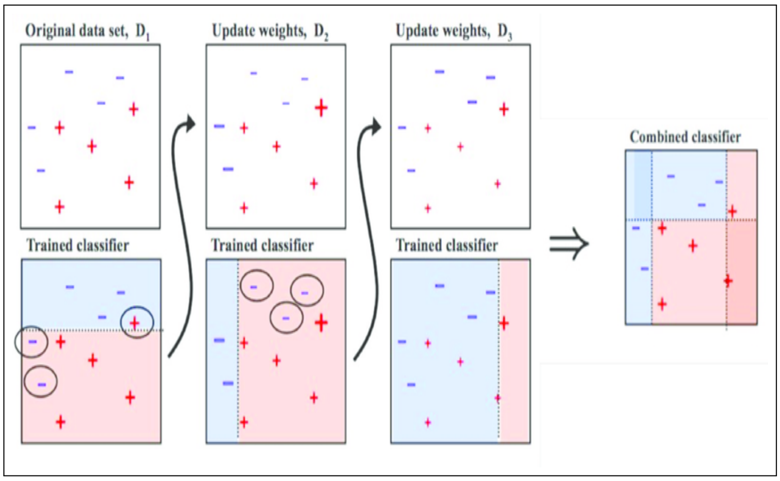

Figure 5.6: AdaBoost Classifier

5.2 Dataset Collection

Description:

The dataset is created in the Google Earth code Editor by adding the Sentinel 2A satellite layer to the world map. Sentinel-2, a Copernicus Earth observation satellite, regularly gathers multi-spectral photos with a high spatial resolution (10 m to 60 m) across land and coastal waterways. Copernicus, an Open Access Hub, offers full access to Sentinel-1, Sentinel-2, Sentinel-3, and Sentinel-5P user products. Previously, it was called Sentinels Scientific Data Hub.We have only considered a tiny section of the Sentinel-2's 13 spectral bands, which range from Visible and Near-Infrared (VNIR) to Shortwave Infrared (SWIR) wavelengths and cover a 290 km orbital swath. By adding a map layer that is filtered by the necessary location and necessary dates on which the sentinel data is available, we may obtain the study area. The sentinel 2 data can be collected from the following link:
https://scihub.copernicus.eu

The figure 5.3 shows the sample images used in this model.

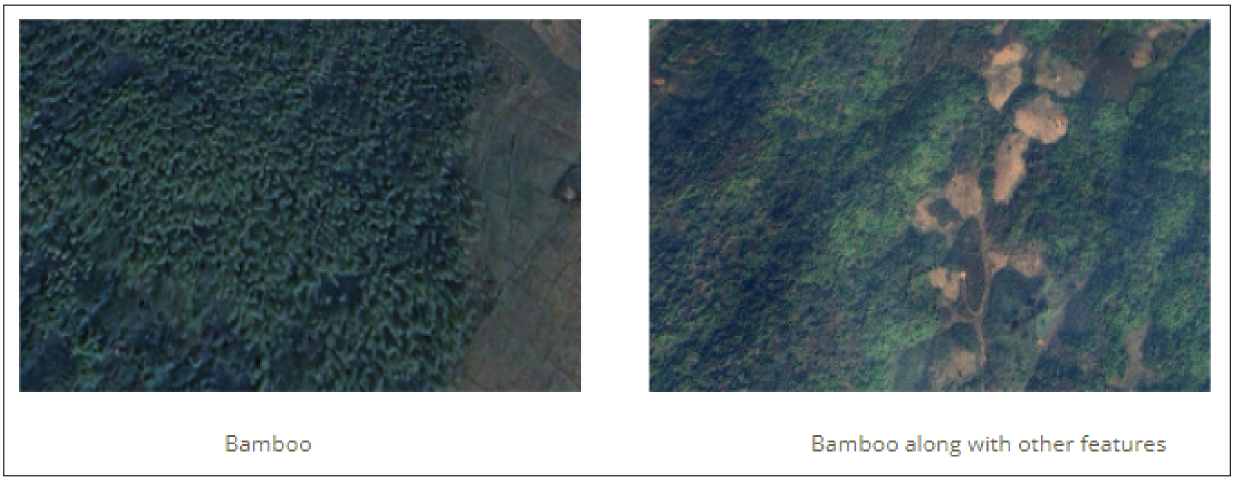

Figure 5.7: Dataset

Chapter 6

IMPLEMENTATION

In this chapter the modules defined for this project are written as algorithms for the better understanding and implementation.

6.1 Data Collection

Input: Shape File and Sentinel 2A data from Copernicus.

Output: Satellite image of the required region.

Methodology:

..

step 1: Import the required district shapefile
step 2: Upgrade Google Earth Code Editor map to import the district layer.
step 3: Import Sentinel data from Copernicus.eu
step 4: Manually annotate the satellite imagery to create point plots
step 5: Convert the annotated data into a CSV file format

6.2 Data Preprocessing

Input: CSV file.

Output: Preprocessed data.

Methodology:

..

step 1: .By using the filling missing values measures like mean,median,mode and also fill them manually.
step 2: Fill the null values present in the data.
step 3: After filling the missing values ,we try to remove the duplicate i.e repeated values from the data.
step 4: For this 'drop_duplicates()' predefined method avaliable in python is used.
step 5: After removing the repeated values ,the data is preprocessed.

6.3 Model Training

Input: Training data, Boosting algorithm that is to be used and the voting algorithms.

Output: Classified model that can predict the results more accuratley.

Methodology:

..

step 1: Determine the sample weight for each sample and assign it.
step 2: To choose the variable to utilise to build the first stump, compute the Gini Index for each variable.
step 3: Identify the Amount of Say the weights of the samples that were erroneously classified add up to the total error.
step 4: Determine the updated sample weights for the following stump.
step 5: Using the revised sample weights, create a bootstrapped dataset that includes the probabilities of each sample being selected.
step 6: The procedure should be repeated n times.

6.4 Performance Evaluation

Input: Classified model.

Output: performace of classified model is evaluated.

Methodology:

..

step 1: The dataset should be split into a training set and a test set as the first step.
step 2: Compare the predicted outputs with the actual labels in the test set.
step 3: Determine the accuracy, which is the percentage of accurate predictions made relative to all predictions made.
step 4: The following formula can be used to calculate accuracy:
Accuracy = Number of correct predictions / Total number of predictions present
step 5: Determine the Recall Value:
Recall is equal to TP / (TP + FN)
step 6: Analyze the accuracy value and Recall Value obtained. A higher accuracy indicates better performance, as it represents the proportion of correct predictions.

Chapter 7

RESULTS

This chapter includes the results of the proposed system

7.1 Data Collection

After selecting the area that is to be classified, a satellite image layer is added to the map of the specified area. Figure 7.1 shows the sentinel images of the three study areas which will be added as a layer on the map.

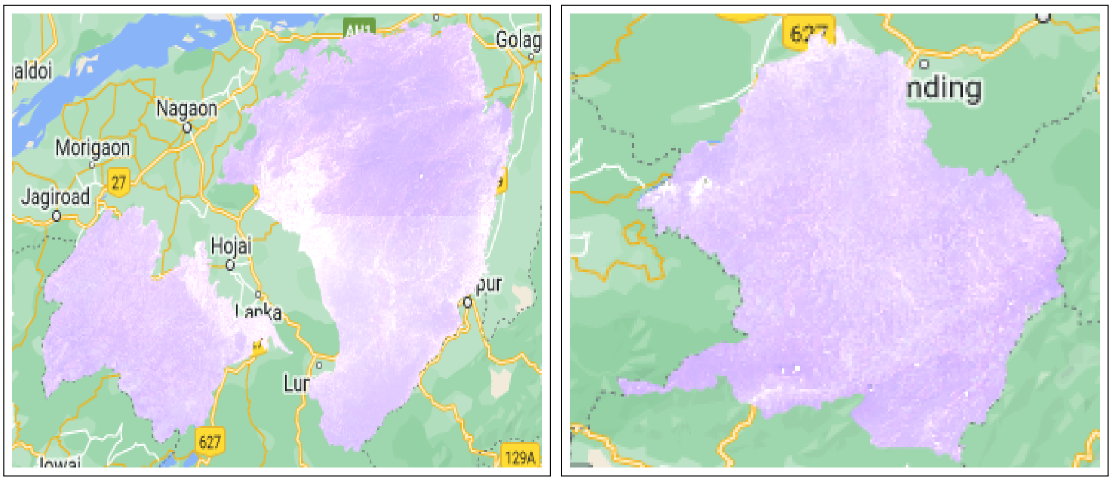

Figure 7.1: Sentinel images of two study areas (Karbi Anglong and Dima Hasao respectively)

7.2 Data Preprocessing

Data points are placed in Karbi Anglong. Figure 7.2 represents the image of Karbi Anglong with 258 data points, 50 for the features like bare soil, bamboo and pure bamboo, 15 points for the feature water bodies and 60 for the features forest and fallow land.

Silver points represents bare soil, light green color points represents pure bamboo, pink points are for bamboo, blue points are for water bodies, dark green points for forest lands and purple points for fallow lands.

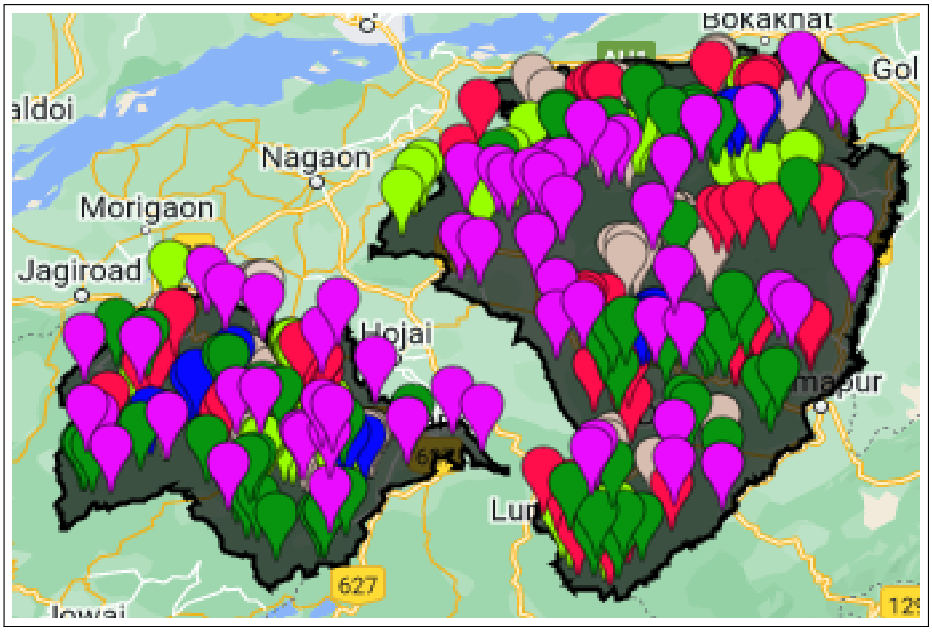

Figure 7.2: Data Points placed on the map in Karbi Anglong Districts.

7.3 Model Training

Four supervised learning algorithms namely SVM, CART, Random Forest and Naive Bayes are trained using Adaboost Technique such that four boosting ensembles are obtrained. Voting algorithms are performed on the four adaboost ensembles. Finally model is trained in a way that it generates optimized output and it can be evaluated in terms of accuracy.

Figure 7.3 shows the buttons present in the GUI

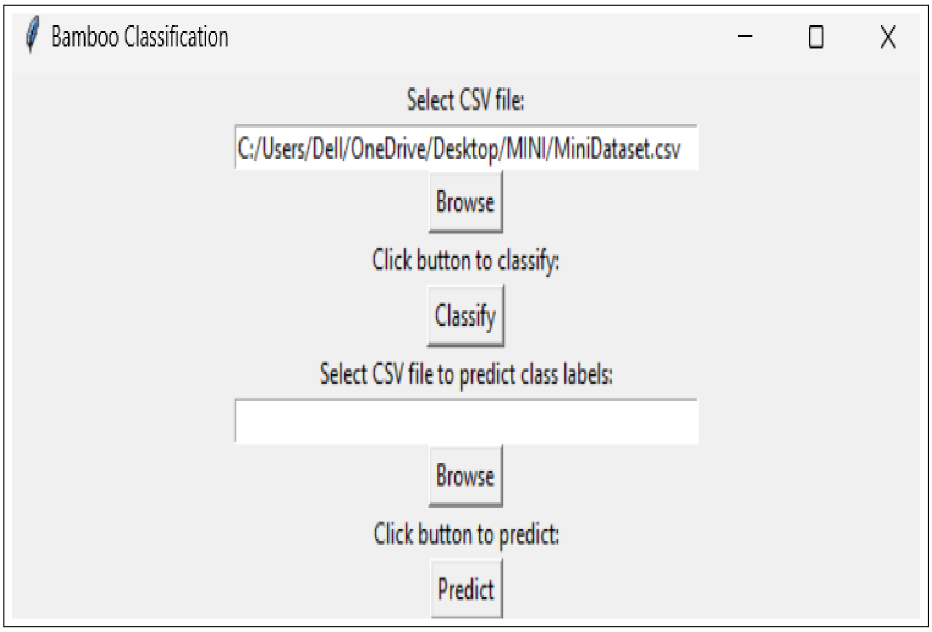

Figure 7.3: buttons present in GUI

In VS Code editor, using python and tkinter library, few buttons were created which allows users to browse the file on which classification needs to be done. The csv file is generated by using points which were plotted earlier in the data collection phase. Next classify button is present to show the results of classification. Once this is done, user will also be allowed to browse another file to predict which type of area is the given data.

7.4 Performance Evaluation

Voting algorithm is applied on the four adaboost ensembles such that we obtain soft, hard, weighted voting applied adaboost ensembles. The performance of these is evaluated using performance evaluation metrics such as Accuracy and Recall Value. The model with high accuracy gives us the best performed model. Here figure 7.4 shows the accuracy and recall for the model. Which implies the highest obtained accuracy as of the soft voting classifier which is 98.53. The other voting algorithms gave outputs which are relatively less compared to soft voting.

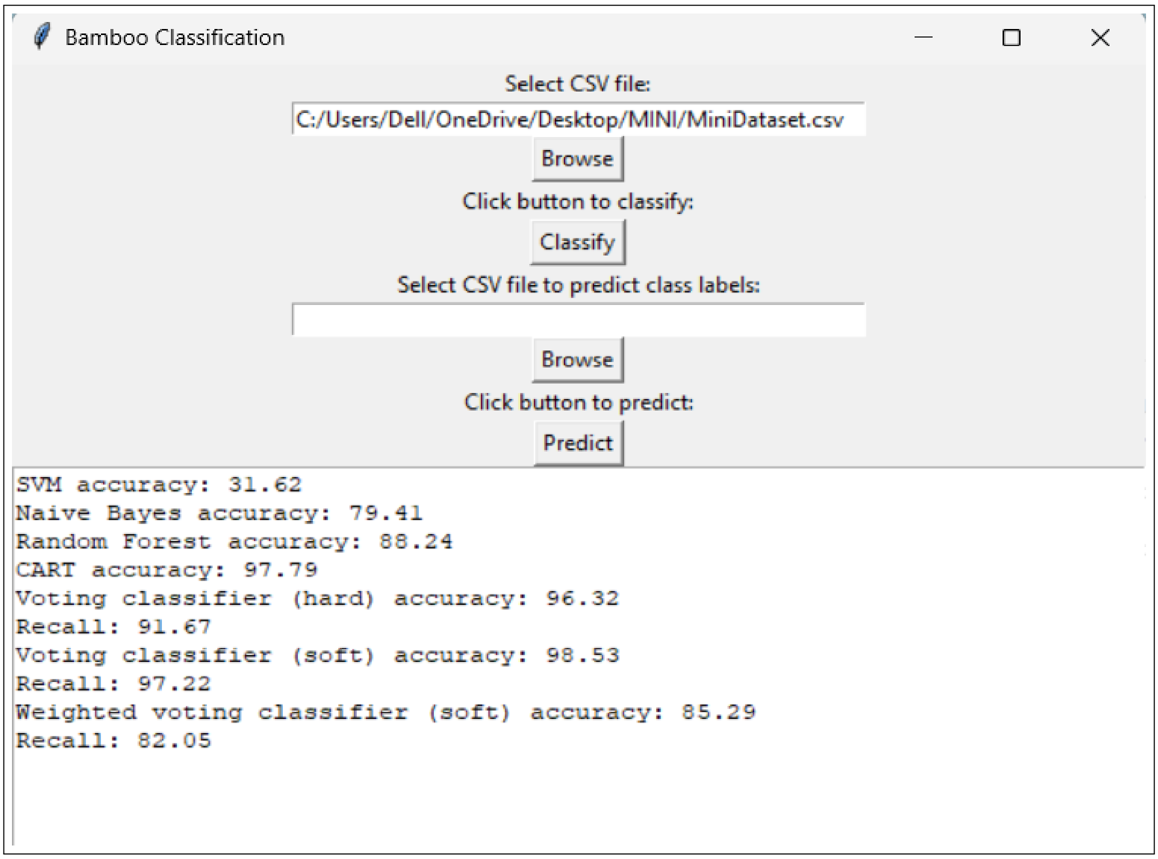

Figure 7.4: Accuracy and Recall

The figure 7.5 shows the predicted class label for unknown tuples

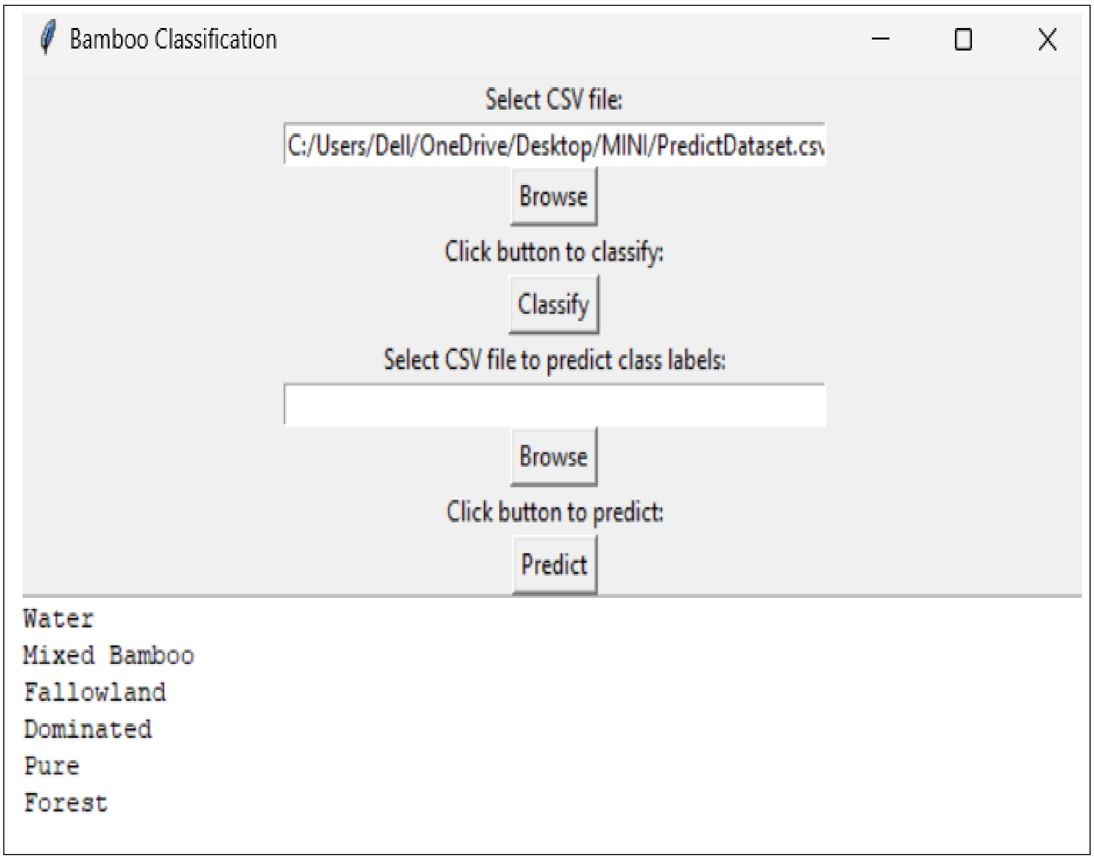

Figure 7.5: Predicting the class label

Chapter 8

CONCLUSION AND FUTURE WORK

This chapter includes the conclusion of the project and future work

In this paper, AdaBoost Ensemble Technique is applied on the supervised learning methods such as Naive Bayes (NB), Support Vector Machine (SVM), Classification and Regression Trees (CART), and Random Forest (RF) to categorize the bamboo region of Karbi Anglong, the district of Assam. These models are tested with validation district Dima Hasao of Assam, India. Datasets with different time stamps were collected from the Sentinel-2 and uploaded into Google Earth code editor platform. In this ensemble learning four base models NB, RF, SVM, CART were considered. And by using boosting technique these models were trained. After training, voting algorithms like Hard,Soft and Weighted voting were applied to combine the results of these trained models.

Future work will combine other Ensemble techniques like Bagging, Stacking and other voting algorithms like Approve, Borda ,Ranked choice and Range voting to overcome the limitations of AdaBoost Ensemble Technique and Hard,Soft,Weighted voting Algorithms.

REFERENCES

[1] Ippolito, P. P. (2021b, December 10). SVM: Feature Selection and Kernels - Towards Data Science. Medium. https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c Last accessed on 22 August 2019

[2] Soni, D. (2019, July 16). What is Bayes Rule? - Towards Data Science. Medium. https://towardsdatascience.com/what-is-bayes-rule-bb6598d8a2fd Last accessed on 19 February 2020

[3] Machine Learning Random Forest Algorithm - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/machine-learning-random-forest-algorithm Last accessed on 4 May 2021

[4] Ghoshal,A.(2023).Bagging and Boosting .EDUCBA.https://www.educba.com/bagging-and-boosting/Last accessed on 13 November 2018

[5] Shabgahi, G. L. (1999). Adaptive majority voter: a novel voting algorithm for real-time fault-tolerant control systems. https://www.semanticscholar.org/paper/Adaptive-majority-voter%3A-a-novel-voting-algorithm-ShabgahiBennett/190efb71da404e4b99cd714ceb08459e680d1276 Last accessed on 5 April 2016

[6] N. C. Yang and H. Ismail, "Voting-Based Ensemble Learning Algorithm for Fault Detection in Photovoltaic Systems under Different Weather Conditions," Mathematics, vol. 10, no. 2, article no. 285, Jan. 2022, doi: 10.3390/math10020285.[Online]. Available: https://www.mdpi.com/2227-7390/10/2/285 Last accessed on 15 January 2022.

[7] L. Wen and M. Hughes, "Coastal Wetland Mapping Using Ensemble Learning Algorithms: A Comparative Study of Bagging, Boosting and Stacking Techniques," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 10, pp. 1683-1696, 2020, doi: 10.1109/JSTARS.2020.3007894.[Online]. Available:https://www.mdpi.com/2072-4292/12/10/1683 Last accessed on 16 September 2020.

[8] K. Thiagarajan, M.M. Anandan, A. Stateczny, P.B. Divakarachari, and H.K. Lingappa, "Satellite Image Classification Using a Hierarchical Ensemble Learning and Correlation Coefficient-Based Gravitational Search Algorithm," Remote Sens., vol. 13, no. 21, article no. 4351, Nov. 2021.

doi: 10.3390/rs13214351.[Online]. Available:https://www.mdpi.com/2072-4292/13/21/4351 Last accessed on 17 November 2021

[9] N. Peppes, E. Daskalakis, T. Alexakis, E. Adamopoulou, and K. Demestichas, "Performance of Machine Learning-Based Multi-Model Voting Ensemble Methods for Network Threat Detection in Agriculture 4.0," Sensors, vol. 21, no. 22, article no. 7475, Nov. 2021, doi: 10.3390/s21227475.[Online]. Available:https://www.mdpi.com/1424-8220/21/22/7475 Last accessed on 16 September 2022.

[10] D. L. N. Hoang, S. Ajibade, J. Dayupay, and O. Oyebode, "Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students," Guangdianzi Jiguang/Journal of Optoelectronics Laser, vol. 41, no. 6, pp. 48-54, Jun. 2022. doi: 10.14113/j.issn.1002-682X.2022.06.008.[Online]. Available: https://www.researchgate.net/publication/361101272 Last accessed on 17 June 2022.

[11] Deepan Perumal and Sudha L.R, "Scene Classification of Remotely Sensed Images using Ensembled Machine Learning Models," in Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication, Proceedings of MDCWC, May 2021, pp. 535-550, doi: 10.1007/978-981-16-0289-4.[Online].Available:https://www.researchgate.net/publication/351964697 Last accessed on 16 May 2021.

[12] F. Leon, S.-A. Floria, and C. Badica, "Evaluating the effect of voting methods on ensemble-based classification," (2019) IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA), Gdynia, Poland, 2019, pp. 25-30, doi: 10.1109/INISTA.2017.8001122. [Online]. Available:https://www.researchgate.net/publication/318974297 Last accessed on 16 July 2020.

[13] R. Harine Rajashree and M. Hariharan. "A Study on Ensemble Methods for Classification." In Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication, Proceedings of MDCWC 2020, pp. 127-136. May 2021. DOI: 10.1007/978-981-16-0289-4_10. [Online]. Available: https://www.researchgate.net/publication/351964892. Last accessed on 21 May 2021

[14] Johnivan, J. (2023). Agile Software Development Methodology Principles. project-management.com. https://project-management.com/agile-software-development-methodologies/ Last accessed on 30 July 2020

[15] Great Learning Team. (2022, November 7). AdaBoost Algorithm: Boosting Algorithm in Machine Learning. Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/adaboost-algorithm/: :text=AdaBoost20algorithm2C20short20for20Adaptive,assigned %20to%20incorrectly%20classified %20instances. Last accessed on 20 April 2021

[16] Sitthiyot, T., Holasut, K. (2020). A simple method for measuring inequality. Palgrave Communications, 6(1). https://doi.org/10.1057/s41599-020-0484-6 Last accessed on 19 June 2022 Saini, A. (2023). Master the AdaBoost Algorithm: Guide to Implementing Understanding AdaBoost. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/: :text=From%20the%20graph%20above%2C20 we,the20classifier%20will%20be%200. Last accessed on 3 May 2020

[17] Saini, A. (2023b). Master the AdaBoost Algorithm: Guide to Implementing Understanding AdaBoost. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners Last accessed on 23 March 2021

[18] Shin, T. (2021, December 15). A Mathematical Explanation of AdaBoost in 5 Minutes. Medium. https://towardsdatascience.com/a-mathematical-explanation-of-adaboost-4b0c20ce4382 Last accessed on 7 August 2022

[19] Alto, V. (2021, December 13). Understanding AdaBoost for Decision Tree - Towards Data Science. Medium. https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851 Last accessed on 19 October 2022

Appendix A

8.1 Appendix A: REPORT PLAGIARISM

ORIGINALITY REPORT

16%	14%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	3%
2	www.mdpi.com Internet Source	1%
3	www.researchgate.net Internet Source	1%
4	gen-live.sei-international.org Internet Source	1%
5	onlineresource.ucsy.edu.mm Internet Source	1%
6	royallib.com Internet Source	1%
7	Kowsalya Thiagarajan, Mukunthan Manapakkam Anandan, Andrzej Stateczny, Parameshachari Bidare Divakarachari et al. "Satellite Image Classification Using a Hierarchical Ensemble Learning and Correlation Coefficient-Based Gravitational Search Algorithm", Remote Sensing, 2021 Publication	1%

8	Submitted to University of Sunderland Student Paper	<1 %
9	Halil Ibrahim Uckol, Taylan Ozgur Bilgic, Suat Ilhan. "Comparative Investigation of Corona Pulse Characteristics under DC and AC Voltages", 2022 IEEE International Conference on High Voltage Engineering and Applications (ICHVE), 2022 Publication	<1 %
10	Submitted to Rochester Institute of Technology Student Paper	<1 %
11	"Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication", Springer Science and Business Media LLC, 2021 Publication	<1 %
12	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
13	docplayer.net Internet Source	<1 %
14	Nikolaos Peppes, Emmanouil Daskalakis, Theodoros Alexakis, Evgenia Adamopoulou, Konstantinos Demestichas. "Performance of Machine Learning-Based Multi-Model Voting Ensemble Methods for Network Threat Detection in Agriculture 4.0", Sensors, 2021	<1 %

Publication

15	www.usgs.gov Internet Source	<1 %
16	Submitted to MCI Management Centre Innsbruck Student Paper	<1 %
17	stackoverflow.com Internet Source	<1 %
18	Submitted to Universiti Teknologi Petronas Student Paper	<1 %
19	guides.library.brandeis.edu Internet Source	<1 %
20	ntnuopen.ntnu.no Internet Source	<1 %
21	Liguo Wang, Yin Shoulin, Hashem Alyami, Asif Ali Laghari, Mamoon Rashid, Jasem Almotiri, Hasan J. Alyamani, Fahad Alturise. "A novel deep learning - based single shot multibox detector model for object detection in optical remote sensing images", Geoscience Data Journal, 2022 Publication	<1 %
22	harshp.com Internet Source	<1 %
23	Medisetty Likhitha, G Kalyani, Tulluri Naga Vennela, Dharmapuri Mahith Paul.	<1 %

"Developing a Pre-Consultation System using Machine Learning for Medical Diagnostics", 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), 2023
Publication

24	ceur-ws.org Internet Source	<1 %
25	cse.anits.edu.in Internet Source	<1 %
26	ethesis.nitrkl.ac.in Internet Source	<1 %
27	"New Metropolitan Perspectives", Springer Science and Business Media LLC, 2021 Publication	<1 %
28	Kalaivani K, M Priya, P Deepan, L.R. Sudha, J Ganesh. "Heart Disease Prediction System Based on Multiple Feature Selection Algorithm with Ensemble Classifier", ECS Transactions, 2022 Publication	<1 %
29	dpnm.postech.ac.kr Internet Source	<1 %
30	etd.aau.edu.et Internet Source	<1 %
31	kirubarajan.nyc3.digitaloceanspaces.com Internet Source	

			<1 %
32	nakug.com	Internet Source	<1 %
33	www.studymode.com	Internet Source	<1 %
34	www.vrsiddhartha.ac.in	Internet Source	<1 %
35	ujcontent.uj.ac.za	Internet Source	<1 %
36	assets.researchsquare.com	Internet Source	<1 %
37	eprints.upj.ac.id	Internet Source	<1 %
38	studentsrepo.um.edu.my	Internet Source	<1 %
39	ijiemr.org	Internet Source	<1 %
40	sportdocbox.com	Internet Source	<1 %
41	vtechworks.lib.vt.edu	Internet Source	<1 %
42	www.application-systems.co.uk	Internet Source	<1 %

43	Submitted to Southampton Solent University Student Paper	<1 %
44	Submitted to University of Greenwich Student Paper	<1 %
45	epdf.pub Internet Source	<1 %
46	silo.tips Internet Source	<1 %
47	academic-accelerator.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

8.2 Appendix B: CODE PLAGIARISM

ORIGINALITY REPORT

32%	20%	23%	30%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Edinburgh Student Paper	7%
2	Submitted to Coventry University Student Paper	5%
3	Submitted to UT, Dallas Student Paper	5%
4	Submitted to Cardiff University Student Paper	4%
5	www.itdaan.com Internet Source	2%
6	Submitted to SUNY, Binghamton Student Paper	2%
7	Sunil Sharma, Lokesh Tharani. "Photonic crystal fiber based automated system to diagnose silent killer", Material Science & Engineering International Journal, 2023 Publication	2%
8	repozitorij.unin.hr Internet Source	2%

9	stackoverflow.com Internet Source	1 %
10	Submitted to Queensland University of Technology Student Paper	1 %
11	Kundu, Bipasha. "Classification and Feature Extraction of Different Hand Movements from the EMG Signal Using Machine Leaning based Algorithms", University of Minnesota, 2021 Publication	1 %
12	www.ijiemr.org Internet Source	1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		