# Comparison of Decision Tree Algorithm and Support Vector Machine for Efficient Soil Data Classification in Terms of Accuracy

**M. Reddaiah[1], R.Bhavani[2*]**

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Saveetha University, Chennai, Tamil nadu. India. Pincode: 602105.

[2*]Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India,Pincode: 602105.

**ABSTRACT**

Aim: The proposed work aims to classify the soil data using the support vector machine algorithm and compare its performance with the decision tree algorithm. Materials and Methods: The study setting of the proposed work utilizes two groups The group 1 is a Support Vector Machine (SVM) algorithm and group 2 is a Decision Tree (DT) algorithm. The sample size for each group is measured as 20 using Gpower of 80%. Results: The experimental results show that the Decision tree algorithm has T-test accuracy (65.95) which is less compared with the Support vector algorithm (76.47). There exists no statistically significant difference among the groups with (P=0.521, >0.05). Conclusion: The obtained results show that the SVM algorithm is performing better than the Decision Tree in terms of accuracy.

**Keywords:** Support Vector Machine, Novel Classification, Decision Tree, Land Cover, Machine Learning, Agriculture, Soil

## INTRODUCTION

Diverse Data Mining techniques can be utilized to determine the fertility of soil and identify the best farmland, especially for agriculture (Jain et al. 2017). Further the classification algorithms are applied to organize knowledge about soil in that area. It determines the properties of soil in the area. Langkawi has a best geological heritage of high value based on its great geological landscape and other features such as the fossils, and erosional effects (Shastry, A, and H 2014). The Langkawi islands are primarily protected under the jurisdiction of the Geoforest Park that are supervised by the Forestry Department. Statistical methods such as the minimum distance and maximum likelihood classifiers have been widely used for the classification of soil data. Soils are classified into group data (Vamanan and Ramar 2011). The applications of Soil classification helps farmers to predict the yield of crops and also determine the best crop which suits the soil(Pruthviraj et al. 2022).

Around 150 IEEE Research papers are identified related to the soil data that have shown that techniques such as evidential reasoning, neural networks, decision trees and Support Vector Machines (SVM) may often be able to classify a data set to a higher accuracy than T-test statistical classifiers. With these techniques, (Pruthviraj et al. 2022) determined that

there are several soil varieties in the world. To predict the type of crop that can be cultivated in that particular soil type it is needed to understand the features and characteristics of the soil type. The authors proposed a method for performing the classification of soil according to the macronutrients and micronutrients. Data mining and machine learning is still an emerging technique in the field of agriculture and horticulture. (Mucherino, Papajorgji, and Pardalos 2009) has done the soil classification according to the soil nutrients which is much beneficial for the farmers to predict which crop can be cultivated in a particular soil type. (Vamanan and Ramar 2011) performed the assessment of several classification techniques and applied them to the database of soil and tested whether there exists a meaningful relationship. The classification done by (Srunitha and Padmavathi 2016) incorporates steps like acquisition of image, preprocessing, extracting features and performing classification. Texture features in images of soil have been extracted utilizing the low pass and Gabor filter. The statistical parameters taken are Mean amplitude, HSV histogram and Standard deviation. (Su et al. 2014) performed the comparative analysis of data mining techniques applied on agricultural fields with concrete examples. (Gholap et al. 2012) has done classification of soil and identified the swelling potential of the clay in Jababeka and Lippo Cikarang and also determined the percentage of mineral content.

(Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021; Bhanu Teja et al. 2021; Veerasimman et al. 2021; Baskar et al. 2021)

The research gap identified from the survey is that the existing methodologies attain less accuracy in classification of soil. The aim of the proposed methodology is to utilize the SVM approach for novel classification and improve the classification accuracy.

## MATERIALS AND METHODS

The research work was performed in the Department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS. Two groups were utilized in this work to perform soil classification. Group 1 is the decision tree algorithm and Group 2 is SVM algorithm. The sample size is estimated using Gpower of 80%. The platform used to evaluate the algorithms was Jupyter (Anaconda) software. The hardware configurations were an Intel core i5 processor with a ram size of 8GB. The Software Configuration of the system is 64-bit, Windows OS, 64 bit processor with HDD of 1TB. The data collection is taken from the open source that is used for software effort estimation using Decision tree algorithm and support vector algorithm Several data used consist of the System data, topography map with 1:50 000 scale and ground truth data. The data of Langkawi Island acquired on March 7, 2005 was used for data classification. They were already geometrically corrected and registered to WGS 84 datum and UTM Zone 47 projection (Vamanan and Ramar 2011).

### Decision tree Algorithm

Given the set of attributes and groups the decision tree frames the set of rules that can be utilized to characterize soil data. The main advantage of the algorithm is

that it is easy to interpret and needs only minimum processing. It can oblige both the numerical and categorical data. It performs the estimation of dependent variables based on the available data. The tree is created based on the attribute values of training data. The data is categorized utilizing the function of data that have gained knowledge. The principle of pruning is established using the tolerance value. NDVI is an index calculated from reflectance measured in the red visible and near infrared channels. Firstly, an NDVI map was produced. From the NDVI map, the NDVI range of each land cover type was generated . The equation of NDVI is shown as below Equation (1).

$$NDVI = (NIR-Red)/(NIR + Red) \quad (1)$$

The purpose of image classification is to group together pixels that have similar patterns of brightness values across a series of image bands or information channels. Therefore, brightness value (BV) for each land cover.

**Pseudocode**

Step 1: Load the NDVI dataset into a variable and check for outliers
Step 2: Outliers are detected using quartile functions
Step 3: Calculated from reflectance measured in the red visible and near infrared channel
Step 4: The range of each land cover types was generated
Step 5: The test results are predicted using decision trees and these are cross validated
Step 6: Accuracy is achieved through means of all decision trees
Step 7: End

**Support vector algorithm**

The classification, training sites representative of land-cover classes of interest were acquired using the ROI tool. The areas selected to serve as training sites should be relatively homogeneous and extensive enough to provide good statistics. The soil images were classified automatically using four kernel types. The kernel types are linear, polynomial, radial basis function and sigmoid. Firstly, the classification was executed by using the default parameters. The parameters are penalty parameter, pyramid levels and classification probability threshold. The default value of the penalty parameter is 100, pyramid levels are 0 and classification probability threshold is 0.

**Pseudocode**

Step 1: Initialize the model with a random value
Step 2: This can also be the average value or mid value of the total values.
Step 3: land cover classes of interest was acquired using ROI tool
Step 4: The new values of penalty parameter were 200,300,and400
Step 5:The new values of classification probability threshold were 0,3,0,7 and 1.
Step 6: The optimum parameter was based on the accuracy assessment
Step 7: End

**Statistical Analysis**

In the current Study the Statistical tool called IBM SPSS is used (Waheed et al. 2006). Using this software's descriptive and group statistics for the accuracy values are calculated. Independent sample tests are taken and significance values are calculated for novel classification. Independent variables are distinct

attributes that are helpful in prediction and dependent variables are improved accuracy values (Gholap et al. 2012).

## RESULTS

The novel Classification of soil is done and the accuracy of soil data is estimated by test set. Table 1 gives the comparison of accuracy between Support vector machine and Decision tree algorithm. Accuracy varies for different iterations for various test set sizes. Table 2 gives the independent sample T test. Support vector machine achieved a mean 76.47, Standard Deviation of 3.516 and the Standard error mean of 1.112. Decision tree achieved a mean 65.95, Standard Deviation of 3.019 and the Standard error mean of .955. Table 3 gives the significance and standard error determination. It is considered to be statistically significant and 95% confidence intervals were calculated. Fig. 1 represents the mean accuracy of the software effort estimation for SVM and Decision tree algorithm. The SVM algorithm obtained 76.47% accuracy and the Decision tree algorithm obtained 65.95% accuracy. The SVM algorithm technique achieved better performance than the Decision tree algorithm.

## DISCUSSION

Experiments were conducted among the study group Decision Tree (DT) and Support Vector Machine (SVM) algorithms by varying sample size for performing novel classification of soil data. From the experiments, it is observed that the proposed SVM performed better in terms of classification of soil data by achieving better accuracy and less error rate compared to the Decision tree algorithm.

Many similar findings prove SVM Algorithm classification is best. (Foody and Mathur 2004) has explored the classification of crops based on ancillary information on soil type and proved to get good accuracy. SVM is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well, it's best suited for classification. (Yi-yang and Nan-ping 2009; Wijaya and Gloaguen 2007) proposed the SVM algorithm retrieval by taking soil dataset and accuracy is higher than 94%. (Pruthviraj et al. 2022) proposed a model that SVM shows higher accuracy in fertilizer recommendation than other algorithms. SVM does not perform very well (Myagila and Kilavo 2021) in the work proposed for detecting sign language. CNN had higher accuracy than SVM.

Although the proposed algorithm yields better accuracy it has certain limitations. Support Vector Machines (SVM) cannot be executed very well when the data set has more sound target classes that are overlapping.and analyze the live data which is constantly changing. In the future various applications can be made by working together with the combination of other machine learning algorithms.

## CONCLUSION

The proposed work has implemented classification of soil data using two different algorithms namely Support Vector Machine Algorithm and Decision Tree Algorithm. Support vector algorithm shows higher accuracy rate and performed better at a more significant rate than that of the decision tree.

## DECLARATIONS

### Conflicts of Interests

No conflicts of interests in the manuscript.

### Authors Contribution

Author MR was involved in data collection, data analysis, and manuscript writing. Author KA was involved in conceptualization, data validation and critical review of manuscript.

### Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

**Funding:** We thank the following organizations for providing financial support that enabled us to complete the study.

1. Alpha Tech Academy
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

## REFERENCES

1. Baskar, M., R. Renuka Devi, J. Ramkumar, P. Kalyanasundaram, M. Suchithra, and B. Amutha. 2021. “Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems.” *Neural Processing Letters*, January. https://doi.org/10.1007/s11063-020-10407-4.
2. Bhanu Teja, N., Yuvarajan Devarajan, Ruby Mishra, S. Sivasaravanan, and D. Thanikaivel Murugan. 2021. “Detailed Analysis on Sterculia Foetida Kernel Oil as Renewable Fuel in Compression Ignition Engine.” *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01328-w.
3. Bhavikatti, Shaeesta Khaleelahmed, Mohmed Isaqali Karobari, Siti Lailatul Akmar Zainuddin, Anand Marya, Sameer J. Nadaf, Vijay J. Sawant, Sandeep B. Patil, Adith Venugopal, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. “Investigating the Antioxidant and Cytocompatibility of Mimusops Elengi Linn Extract over Human Gingival Fibroblast Cells.” *International Journal of Environmental Research and Public Health* 18 (13). https://doi.org/10.3390/ijerph18137162.
4. Foody, Giles M., and Ajay Mathur. 2004. “Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification.” *Remote Sensing of Environment* 93 (1): 107–17.
5. Gholap, Jay, Anurag Ingole, Jayesh Gohil, Shailesh Gargade, and Vahida Attar. 2012. “Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1206.1557.
6. Jain, Nishit, Amit Kumar, Sahil Garud, Vishal Pradhan, and Prajakta Kulkarni. 2017. “Crop Selection Method Based on Various Environmental Factors Using Machine Learning.” *International Research Journal of Engineering and Technology (IRJET)* 4 (02): 551–58.
7. Karobari, Mohmed Isaqali, Syed Nahid Basheer, Fazlur Rahman Sayed, Sufiyan Shaikh, Muhammad Atif

Saleem Agwan, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "An In Vitro Stereomicroscopic Evaluation of Bioactivity between Neo MTA Plus, Pro Root MTA, BIODENTINE & Glass Ionomer Cement Using Dye Penetration Method." *Materials* 14 (12). https://doi.org/10.3390/ma14123159.

8. Karthigadevi, Guruviah, Sivasubramanian Manikandan, Natchimuthu Karmegam, Ramasamy Subbaiya, SivasankaranChozhavendhan, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2021. "Chemico-Nanotreatment Methods for the Removal of Persistent Organic Pollutants and Xenobiotics in Water - A Review." *Bioresource Technology* 324 (March): 124678.
9. Mucherino, Antonio, PetraqPapajorgji, and Panos M. Pardalos. 2009. *Data Mining in Agriculture*. Springer Science & Business Media.
10. Muthukrishnan, Lakshmipathy. 2021. "Nanotechnology for Cleaner Leather Production: A Review." *Environmental Chemistry Letters* 19 (3): 2527–49.
11. Myagila, Kasian, and Hassan Kilavo. 2021. "A Comparative Study on Performance of SVM and CNN in Tanzania Sign Language Translation Using Image Recognition." *Applied Artificial Intelligence*. https://doi.org/10.1080/08839514.2021.2005297.
12. Preethi, K. Auxzilia, K. Auxzilia Preethi, Ganesh Lakshmanan, and Durairaj Sekar. 2021. "Antagomir Technology in the Treatment of Different Types of Cancer." *Epigenomics*. https://doi.org/10.2217/epi-2020-0439.
13. Pruthviraj, G. C. Akshatha, K. Aditya Shastry, Nagaraj, and Nikhil. 2022. "Crop and Fertilizer Recommendation System Based on Soil Classification." In *Recent Advances in Artificial Intelligence and Data Engineering*, 29–40. Springer Singapore.
14. Sawant, Kashmira, Ajinkya M. Pawar, Kulvinder Singh Banga, Ricardo Machado, Mohmed IsaqaliKarobari, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "Dentinal Microcracks after Root Canal Instrumentation Using Instruments Manufactured with Different NiTi Alloys and the SAF System: A Systematic Review." *NATO Advanced Science Institutes Series E: Applied Sciences* 11 (11): 4984.
15. Shanmugam, Vigneshwaran, Rhoda Afriyie Mensah, Michael Försth, Gabriel Sas, Ágoston Restás, Cyrus Addy, Qiang Xu, et al. 2021. "Circular Economy in Biocomposite Development: State-of-the-Art, Challenges and Emerging Trends." *Composites Part C: Open Access* 5 (July): 100138.
16. Shastry, K. Aditya, Sanjay H. A, and Kavya H. 2014. "A Novel Data Mining Approach for Soil Classification." In *2014 9th International Conference on Computer Science Education*, 93–98.
17. Srunitha, K., and S. Padmavathi. 2016. "Performance of SVM Classifier for Image Based Soil Classification." In *2016 International Conference on Signal Processing, Communication, Power and Embedded System*

*(SCOPES)*, 411–15.

18. Su, Xiaoquan, Jianqiang Hu, Shi Huang, and Kang Ning. 2014. "Rapid Comparison and Correlation Analysis among Massive Number of Microbial Community Samples Based on MDV Data Model." *Scientific Reports* 4 (September): 6393.
19. Vamanan, R., and K. Ramar. 2011. "Classification of Agricultural Land Soils a Data Mining Approach." *International Journal on Computer Science and*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.8621&rep=rep1&type=pdf.
20. Veerasimman, Arumugaprabu, Vigneshwaran Shanmugam, Sundarakannan Rajendran, Deepak Joel Johnson, Ajith Subbiah, John Koilpichai, and Uthayakumar Marimuthu. 2021. "Thermal Properties of Natural Fiber Sisal Based Hybrid Composites – A Brief Review." *Journal of Natural Fibers*, January, 1–11.
21. Waheed, T., R. B. Bonnell, S. O. Prasher, and E. Paulet. 2006. "Measuring Performance in Precision Agriculture: CART—A Decision Tree Approach." *Agricultural Water Management* 84 (1): 173–85.
22. Wijaya, Arief, and Richard Gloaguen. 2007. "Comparison of Multisource Data Support Vector Machine Classification for Mapping of Forest Cover." In *2007 IEEE International Geoscience and Remote Sensing Symposium*, 1275–78.
23. Yi-yang, Gao, and Ren Nan-ping. 2009. "Data Mining and Analysis of Our Agriculture Based on the Decision Tree." In *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, 2:134–38.

## TABLES AND FIGURES

**Table 1.** Comparison between SVM and Decision tree. Accuracy, values obtained for SVM and Decision tree algorithms are compared for various datasets

| S. NO | SVM | DECISION TREE |
|---|---|---|
| | Accuracy | Accuracy |
| 1 | 61.56 | 71.44 |
| 2 | 62.59 | 72.98 |
| 3 | 63.29 | 73.34 |
| 4 | 64.56 | 74.56 |
| 5 | 65.05 | 75.15 |
| 6 | 66.48 | 76.56 |
| 7 | 67.64 | 78.96 |

| | | |
|---|---|---|
| 8 | 68.58 | 79.84 |
| 9 | 69.25 | 80.21 |
| 10 | 70.50 | 81.67 |

**Table 2.** T-test statistical analysis of two algorithms are obtained for 10 iterations. Decision Tree has less mean accuracy 65.95 compared to SVM (76.47).

| | **Algorithm** | **N** | **Mean** | **Std.Deviation** | **Std. Error Mean** |
|---|---|---|---|---|---|
| **Accuracy** | SVM | 10 | 76.47 | 3.516 | 1.112 |
| | DECISION TREE | 10 | 65.95 | 3.019 | .955 |

**Table 3.** Independent sample test for significance and standard error determination considering equal and not equal variances and 95% confidence intervals were calculated.

| | **Levene's Test for Equality of Variances** | | **T-test for Equality of Means** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **t** | **df** | **Sig** | **Mean Difference** | **Std.Error Difference** | **95% Confidence interval of the Difference** | |
| | **F** | **Sig.** | | | | | | **Lower** | **Upper** |
| Equal variance | .428 | .521 | -7.179 | 18 | .000 | -10.521 | 1.465 | -13.600 | -7.442 |
| Equal variances not assumed | | | -7.179 | 17.59 | .000 | -10.521 | 1.465 | -13.605 | -7.437 |

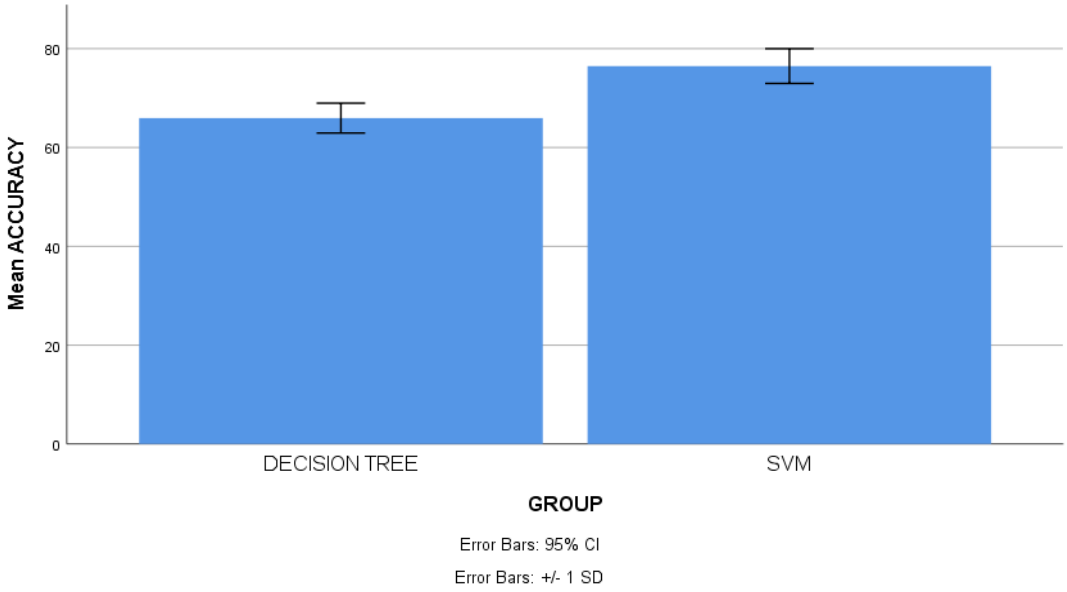


**Fig. 1**. Prediction accuracy for the two algorithms. The accuracy of the SVM algorithm is better than the accuracy of the Decision tree algorithm. X Axis: SVM vs Decision tree algorithm. Y Axis is mean accuracy of detection +/-1 SD.