2021

Seminar Paper

Comparative study on burnt area mapping using Support Vector Machine and Random Forest.

University of Salzburg

Department of Geoinformatics

Prof. Stefan Lang

Name:	Pratichhya Sharma
Matriculation Number:	12031082
Program:	Copernicus Master in Digital Earth
Course:	Analysis and Modeling (Remote Sensing)
Email:	mepratichhya@gmail.com / s1078794@stud.sbg.ac.at
Date of Submission:	22 July 2021

Table of Contents

ABBERVATION 2
ABSTRACT 3
INTRODUCTION 3
DATASET AND STUDY AREA 5
METHODOLOGY 7
DATA PRE-PROCESSING 8
CLASSIFICATION 8
RESULT AND ANALYSIS 9
DISCUSSION 11
CONCLUSION 11
REFERENCES 12

FIGURES:

FIGURE 1: SPECTRAL BANDS OF SENTINEL 2 A 5
FIGURE 2: GROUND TRUTH (SOURCE: HTTPS://DATA.GOV.AU/DATA/DATASET/201920FY-BUSHFIRE-BOUNDARIES) 6
FIGURE 3: TRAINING AND TESTING SET (GENERATED FROM: HTTPS://SCIHUB.COPERNICUS.EU/DHUS/#/HOME) 6
FIGURE 4: INFERENCING SET (GENERATED FROM: HTTPS://SCIHUB.COPERNICUS.EU/DHUS/#/HOME) 7
FIGURE 5: WORKFLOW FOR BURNT AREA MAPPING 7
FIGURE 6: PARAMETER TUNING 8
FIGURE 7: TESTING ACROSS TRAINED AREA 9
FIGURE 8: TESTING ACROSS A NEW AREA 10
FIGURE 9: MODEL EVALUATION 10

ABBERVATION

EO:	Earth Observation
GIS:	Geographic Information System
ML:	Machine Learning
RBF:	Radial Base Function
RF:	Random Forest
RGB:	Red Green Blue
RS:	Remote Sensing
SVM:	Support Vector Machine
NIR:	Near Infra-Red
SWIR:	Short Wave Infra-Red

ABSTRACT

Machine learning has been widely used in environmental science since the early 1990s (Jain et al., 2020). Here, we present a scoping review of two different machine learning algorithms for burnt area mapping in the southern region of Australia. Our objective is to make a comparative study of Support Vector Machine and Random Forest classifiers in burnt area mapping using Sentinel-2 imagery. We first present our findings on the classification of a burnt area using an SVM classifier and its parameter tuning. Then the process is then followed by using RF as a second machine learning algorithm for observing the performance differences between these algorithms. The overall accuracy was ranging from 84% to 90% with different parameters in both models. Among the two ML algorithms, RF with optimized tree number gave the best performance followed by SVM with RBF Kernel. Finally, the advantages and disadvantages of them were discussed, and it was concluded that, despite machine learning models' ability to train in different ways, knowledge in both remote sensing and characteristics of phenomenon is required to provide realistic fire mapping. The study demonstrated that SVM and RF can handle learning tasks with a small training dataset and produce significant results.

Keywords: Sentinel-2; Random Forest (RF); Support Vector Machine (SVM); Burnt area mapping

---

INTRODUCTION

Forest Fire is a natural hazard and can occur due to natural and manmade interactions. In many cases, it is necessary for the health and renewal of the forest and other ecosystems. However, it may result in poor air quality, property damage and adverse human health as well as destroying flora and fauna and its habitat. By 2030, forest fires are expected to have destroyed half of the world's forests (Mahmoud & Ren, 2019). To assess the impact of fire on the forest, precise quantitative and qualitative estimates of burn-area are required (Bar, Parida, & Pandey, 2020) which provides information on their occurrence, propagation and dynamics. As the frequency and intensity of forest fires has increased around the world, research into mapping and detecting forest fires using remote sensing has accelerated because of its capability to obtain large amount of information in multiple spatio-temporal resolution.

Advances in remote sensing (RS) technologies and methods have greatly increased geospatial analysis access to Earth Observations (Sheykhmousa et al., 2020). Remote sensing imagery can aid in the management of wildfires before, during, and after the occurrence. Remote sensing related techniques are used in detecting, mapping, and monitoring of areas affected by forest fires as these areas burned by fire has similar spectral range (Pacheco, Junior, Ruiz-Armenteros, & Henriques, 2021). Analyzing these images aids in knowledge of pre- and post-burn fire conditions. It is not only the appropriateness of imagery or spatial information that affects the analysis results, but also the right choice of classification method that has remarkable influence (Lu & Weng, 2007). There are numerous tools and approaches that can be employed in this process, with machine learning being one of the most extensively employed. As quoted by Arthur Samuel in 1959, "Machine Learning can be defined as a field of the study that provides machine with an ability to perform task based on what that have learned and present the findings without being explicitly programed". It can be seen as a subset of artificial intelligence that uses algorithms to try to

replicate the human brain. Machine learning has been used successfully in the processing and analysis of remote sensing data. Remote sensing has evolved into a multidisciplinary field, with machine learning and signal processing algorithms playing an important role, in order to treat the requested data efficiently and provide accurate products (Camps-Valls, 2009). Supervised Learning is a class of machine learning that learns from labeled data and is widely utilized for remote sensing. This method is preferred when sufficient amount of training dataset is available. In this study, two different machine learning algorithms are implemented namely Support Vector Machine **(SVM)** and Random Forest **(RF)** to perform forest fire mapping using satellite imagery.

SVM is a supervised learning algorithm that divides the data into different classes using a hyperplane (cloudml, 2018). It seeks to identify the locations of decision boundaries that produce the best results. SVM classifiers select from an infinite number of linear decision boundaries when dealing with two-class problems, but when dealing with classes that are not linearly separable, it tries to find the hyperplane that maximizes the margin (Pal & Mather, 2005). This iterative process of building an optimally decided classifier is described as the learning process (Sheykhmousa et al., 2020). Once the hyperplane is determined at certain position, we create a parallel plane and make sure that it passes through the nearest data points. In this classifier, support vectors are data points (one or many) that are closer to the hyperplane and influence the position and orientation of the hyperplane (cloudml, 2018). One of the main advantages of SVM compared to other techniques in ML is its insensitivity to the distribution of underlying data (Sheykhmousa et al., 2020). The ability to use new kernels instead of linear boundaries increases the flexibility of SVMs for decision making, which in turn result for better performance of this algorithm (ibid.). Although SVMs have been widely used for classification problems, recently developed deep learning algorithms have proven to be more efficient than SVMs when given a large amount of training data (Jain et al., 2020). However, it was studies that for problems with limited training samples, SVMs might give better performances in comparison with the deep learning-based classifiers.

Random forest is an ensemble method, that produces a number of decision trees to predict the outcome. Each tree provides a classification, and we say that the tree "votes" for that class. A random selection of feature is evaluated in each tree that defines the title of being random. The strength of individual trees in the forest and their correlation determines the generalization error of a forest of tree classifiers (Breiman, 2001). In simple terms, random forest selects random samples from a set of data and builds a decision tree for each sample based on these samples to obtain a prediction result from each tree. Counting the votes for each predicted outcome will aid in selecting the final prediction with the most votes. This algorithm's high performance is obtained by minimizing tree correlation while reducing model variance, resulting in a large number of different trees providing greater accuracy than individual trees. This improved performance, however, comes at the cost of greater bias and a loss of interpretability (Jain et al., 2020). The final prediction voting in RF mitigates the problem of overfitting with different tree structures and splitting variables (Sheykhmousa et al., 2020). This classification method is highly recommended for dealing with high data dimensions and multicollinearity (Belgiu & Drăgu, 2016). It is also very much popular in the field of remote sensing analysis due to its simplicity methods and effectiveness in performance.

In this paper, we use Sentinel 2 imagery to compare the performance of these two different machine learning algorithms in remote sensing image analysis. Here, we try to highlight the performance of these algorithms for earth observation datasets, as well as their pros and cons. This paper discusses the performance in terms of flexibility and efficiency of machine learning algorithms in remote sensing for burnt mapping. Finally, we'll present our findings and discuss potential applications for these two supervised machine learning algorithms in remote sensing analysis.

DATASET AND STUDY AREA

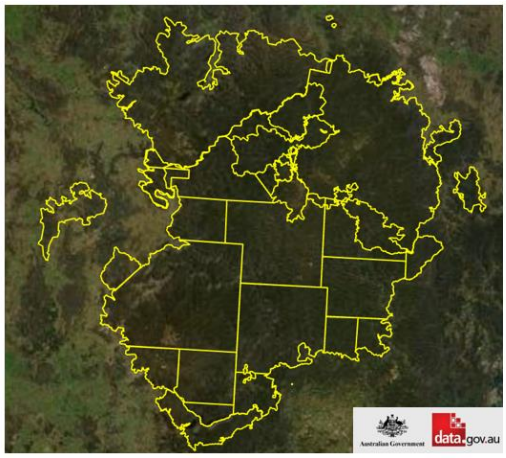
In order to perform this study, a systematic literature review was performed using various web-based bibliographic database from related disciplines based on which selection of dataset was made accordingly. Sentinel 2 imagery with their specific spectral bands were selected and used in mapping the area affected by fire. Spectral bands of Sentinel 2 are as shown in figure 1.

Spectral Band	Center Wavelength (nm)	Band Width (nm)	Spatial Resolution (m)
Band 1	443	20	60
Band 2	490	65	10
Band 3	560	35	10
Band 4	665	30	10
Band 5	705	15	20
Band 6	740	15	20
Band 7	783	20	20
Band 8	842	115	10
Band 8a	865	20	20
Band 9	945	20	60
Band 10	1380	30	60
Band 11	1610	90	20
Band 12	2190	180	20

Figure 1: Spectral Bands of Sentinel 2 A

But only Band 2, Band 3, Band 4, Band 8 and Band 12 were focused on this study because they represent the RGB (visual), near infrared and SWIR because they are sensitive to vegetation, temperature and also the visual colors.

Southern part of Australia was focused as they had lately been in the midst of a major disaster and mainland was ravaged by the biggest bushfires the country had ever seen in late 2019 (Singh, 2020). The burnt area was identified using the ground truth layer for the date of 2020-01-15, provided by the Australian Government as open data. With the idea that the area might not be well captured due to flames, satellite imagery was chosen for 5 days later than the ground truth layer acquisition data i.e., 2020-01-20.



Field	Value
Title	2019-20 Financial Year Bushfire Boundaries
Type	Dataset
Language	English
Licence	Creative Commons Attribution 4.0 International
Data Status	active
Update Frequency	never
Landing Page	https://data.gov.au/data/dataset/f91c2ffd-a0a0-4042-a8ac-71757a6ef727
Date Published	2020-07-01
Date Updated	2020-07-15
Contact Point	National Bushfire Recovery Agency +61 2 6228 6300 reporting@bushfirerecovery.gov.au
Temporal Coverage	2019-07-01 - 2020-06-30
Geospatial Coverage	Australia
Jurisdiction	Commonwealth of Australia
Data Portal	data.gov.au
Publisher/Agency	National Bushfire Recovery Agency
Geospatial Topics	Boundaries

Figure 2: Ground Truth (Source: https://data.gov.au/data/dataset/201920fy-bushfire-boundaries)

The choice of the study area for this region of Australia was solely based on satellite imagery with the least amount of cloud and a ground truth shapefile. Figure 3 shows the subset of the area of interest chosen to be used as training and testing sets in a way that it includes burned and not burned area. The burnt area was clearly depicted in this region along with the variation of different land types. 80% of the entire image was used as training set for the machine learning model, whereas the testing set consisted of 20%.

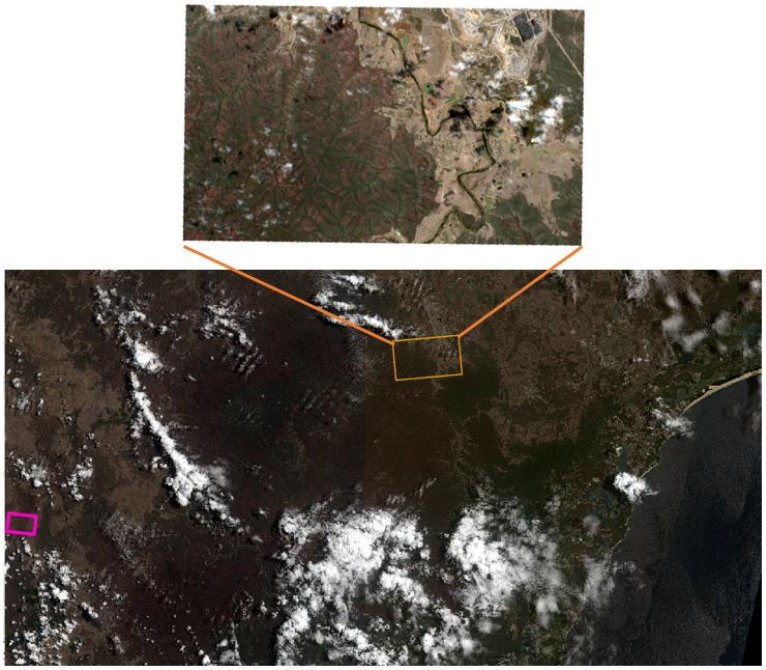


Figure 3: Training and Testing set (Generated from: https://scihub.copernicus.eu/dhus/#/home)

In order to infer and check the capacity of the models that were generated, a new area was chosen as shown in Figure 4. This area had slightly different land cover than the area on which model was trained

but burned area was identified with the help of ground truth. It was selected for visually evaluating and for comparing how well these models work in an unseen environment.

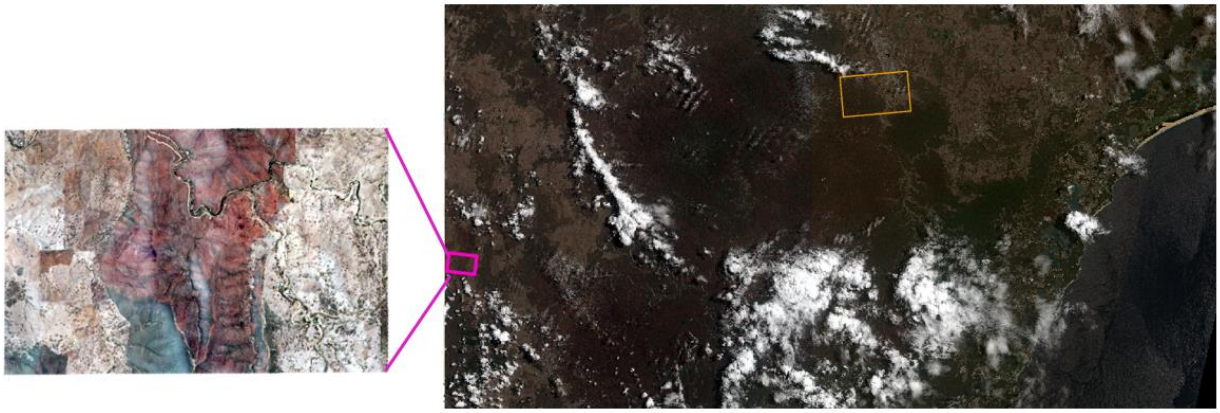


Figure 4: Inferencing set (Generated from: https://scihub.copernicus.eu/dhus/#/home)

METHODOLOGY

This section is dedicated for summarizing entire workflow performed in burned area mapping. It was completed in two main phases: Data Preparation in first phase and Classification using Machine Learning in its second phase. Data preparation included the steps of data acquisition based on the requirement, followed by preprocessing on specific bands of the imagery and ultimately generating random points for training with the help of the ground truth layer. Except for data splitting, the whole workflow for initial phase was completed in ArcGIS Pro software from ESRI, whereas the second phase of image classification into burnt and not burnt was completed entirely using Python programming. Classification of the dataset was done using 2 different supervised machine learning algorithms. Several models were created with the alteration in its parameters for each of SVM and RF. Performance of these models were evaluated with the creation of confusion matrix.

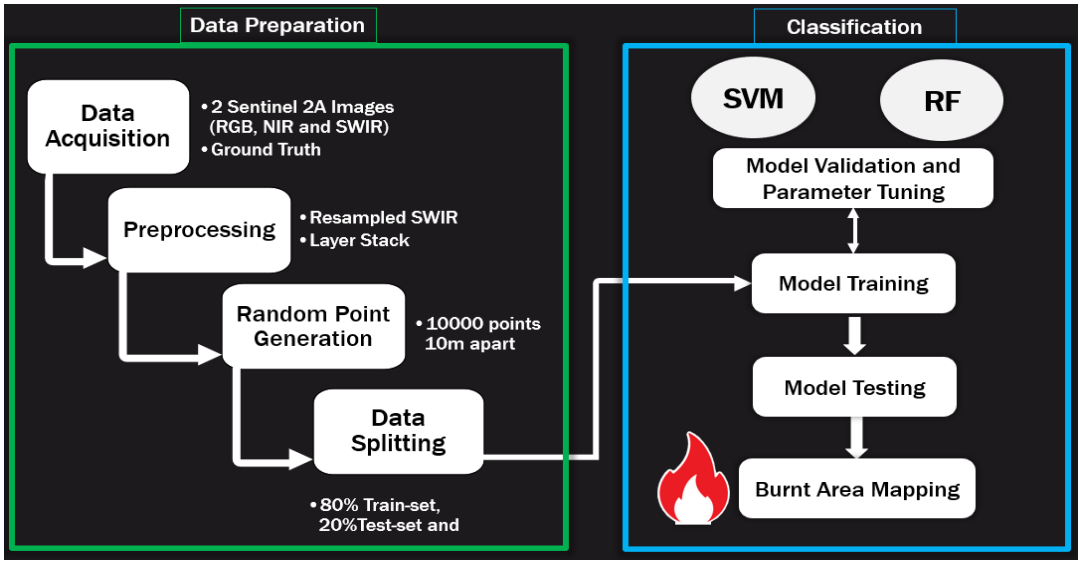


Figure 5: Workflow for burnt area mapping

DATA PRE-PROCESSING

The comparative study started with the acquisition of 2 Sentinel Image covering training, testing and inferencing area of interest and a brunt area ground truth layer. RGB, NIR and SWIR bands of the image were considered in this study realizing their characteristics for visual examination, for vegetation analysis, and for temperature anomaly analysis (Schepers et al., 2014). Preprocessing included the basic steps applied to obtain ready to use raster layer for the area of interest. To maintain data uniformity, the SWIR band was resampled to the same resolution as the RGB band, followed by layer stacking of the previously mentioned 5 bands. With creation random points from ground truth, it was rasterized and reclassified as either burned or not burnt. Then using the Random Point Generator tool in ArcGIS pro 10000 points were generated at 10m apart. The part of splitting this dataset into training and testing set was done in python.

CLASSIFICATION

As a simple working procedure of machine learning, it utilizes traditional classifier techniques that follow five step approaches consisting of: (1) Burnt area identification; (2) feature extraction; (3) feature selection; (4) model training and validation; and (5) feature classification. A pattern data discovered in identification phase is then selected after extraction to use in model creation for each algorithm. With the label samples generated in the pre-processing step, the model is trained then fitted with the optimized parameter and finally applied to the testing and inference set.

Tuned parameters play a significant role in producing good results from a ML models. Tuning was done before testing the results to find the best value for specific parameters. It is generally impossible to determine the best parameters in advance but a hit and trial method can be included to identify the best possible solution (Koehrsen Will, 2018). When using the SVM classifier with radial kernel (RBF), two parameters were tuned: the optimum parameters cost (C) and the kernel width parameter (gamma). The C parameter was for determining the size of error allowing for non-separable training data which make it possible for the model to adjust the rigidity of training data while kernel width affects the smoothing form of the Hyperplane class (Yıldırım Soner, 2020). While implementing RF, only the number of trees was tuned to get the best performance of these models, leaving the rest of its parameters as default. According to several studies, satisfactory results can be achieved using the default parameters. However, a large number of trees will provide a stable result of variable importance, but using more than the required number of trees may be inefficient and unnecessary (Thanh Noi & Kappas, 2017).

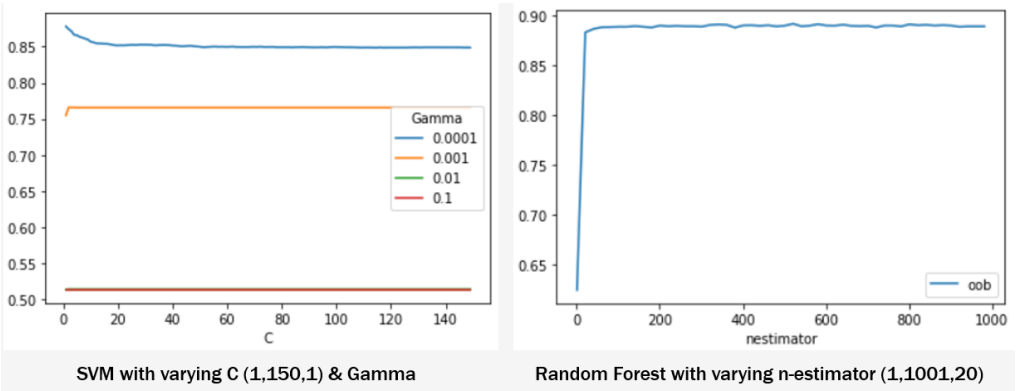


Figure 6: Parameter Tuning

In SVM with RBF kernel, higher overall accuracy was obtained when the value of C was set to 1 and gamma to 0.0001, as shown in figure 6. For this, the model is iterated for 150 values of C and four different gamma values for each C. On the other hand, in the case of RF, though it is based on a random selection of samples and accuracy varied in each compilation, it was evident that several trees set as 521 gave better result while iterating the model through 1 to 1001 with 20 steps.

RESULT AND ANALYSIS

Burnt area mapping with 4 bands of Sentinel 2 imagery using 2 different machine learning algorithms were implemented, evaluated and compared. We explored a variety of tuning parameters for each model to determine the best parameters based on overall classification accuracy. To compare the performance of these models, the classified results under the optimal parameters of each classifier were used in this study. With the completion of the task, performance of each model was compared numerically and visually. Figure 7 and Figure 8 shows the spatial distribution of burned area for Random Forest and Support Vector Machine with variation in parameters. Multiple models were run for SVM and RF with different parameter setting.

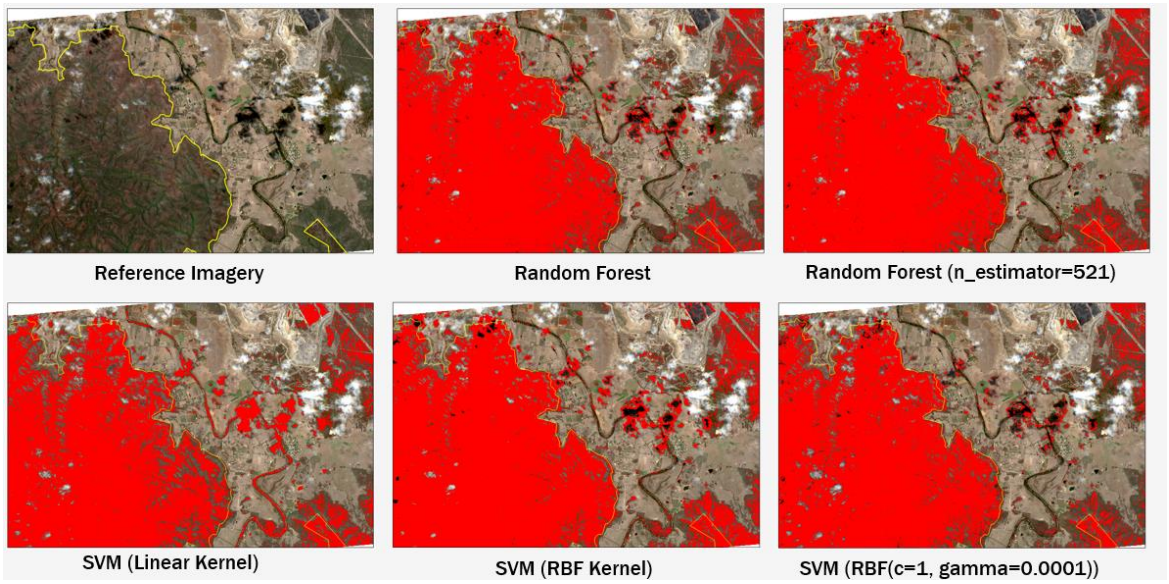


Figure 7: Testing across trained area

Figure 7 represents the area where the model was trained. It shows that in the case of SVM, when the kernel is set to RBF, the results are better than when the kernel is linear. Furthermore, as mentioned in the previous section, the best possible value was determined with the help of parameter tuning. In this case, the results of RF with 521 trees and SVM with RBF kernel were nearly identical. They were able to identify burned areas and classify non-burned areas such as clouds and rivers.

All models previously generated were re-tested over a completely new area as shown in Figure 8.

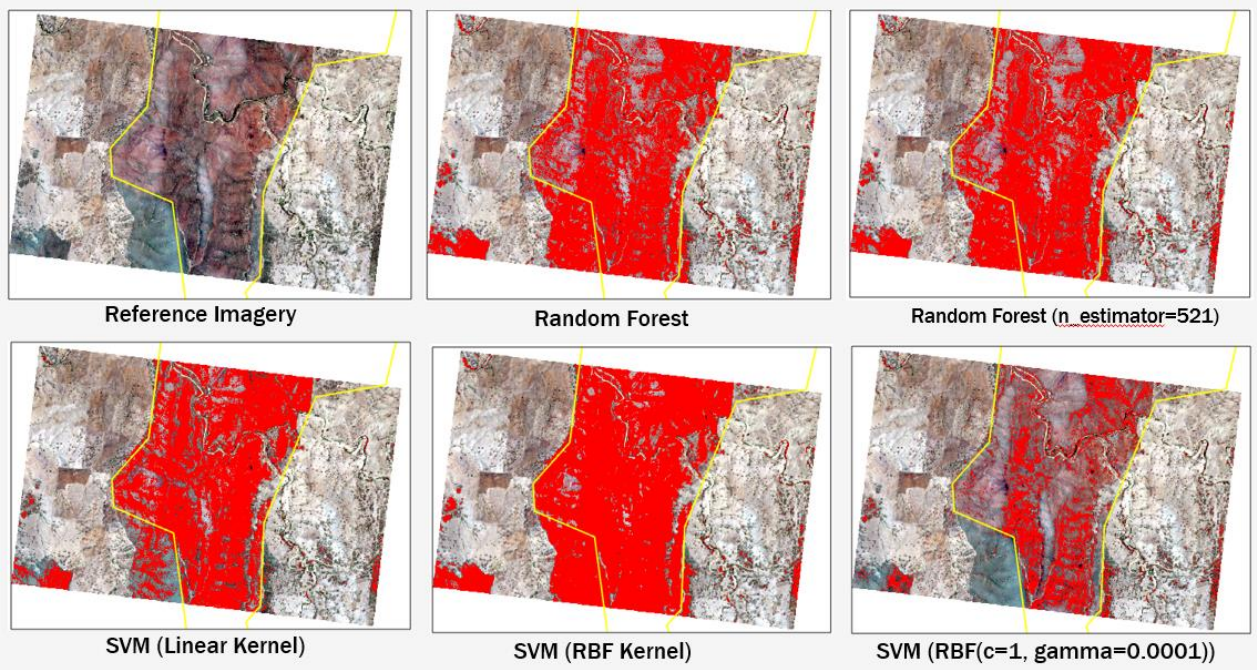


Figure 8: Testing across a new area

The outcome of these models over inferenced area was a bit different from the result obtained over the trained area. In this case, SVM with RBF kernel with tunned parameter seems to have comparatively match to the ground truth than that of RF model. In the RF model, we observed that it was misclassifying all forest areas as burnt areas, which can be argued to be true as the date of acquisition of image is 5 days later than that of ground truth. But the reason could also be due to the overfitting of the model.

The classification accuracy was calculated using validation data and the results of the classification parameter analysis, which included the computation of precision, recall, and overall accuracy.

Method	Parameters	Overall Accuracy	F1 Score	Precision	Recall
SVM (Linear)	Default	84.59%	0.8271	0.8271	0.8581
SVM (RBF)	Default	88.09 %	0.9236	0.9236	0.8196
SVM (RBF)	C=1 Gamma= 0.0001	88.34%	0.8637	0.8637	0.8988
Random Forest	Default	89.34%	0.9011	0.9011	0.8738
Random Forest	n_estimator = 521	89.74 %	0.9087	0.9087	0.8717

Figure 9: Model Evaluation

When compared to the SVM algorithm, the RF algorithm produced the highest classification quality values. It was realized that in order to improve learning algorithm, more training examples are needed.

DISCUSSION

Machine learning techniques have been used in the development, accuracy, computational efficiency, and application of remote sensing analysis in a range of fields. As a result, the benefits of powerful but efficient ML methods for wildfire mapping are widely anticipated. However, each of these algorithms has its own set of benefits and drawbacks. SVM is useful for classification problems and can model extremely complex dimensions, but it is difficult to understand. This algorithm is memory intensive and does not provide probability estimates (cloudml, 2018). They also use less memory because they only use a subset of training points in the decision phase, and they work well with a large dimensional space and a clear margin of separation (Datacamp, 2019). Because of its long training time, this algorithm is not suitable for large datasets. Random Forest is well-known for its high accuracy and has an automatic feature selection technique that determines the most important features. It is capable of handling missing data but this algorithm has the disadvantage of giving user very little control over what goes on inside the algorithm (cloudml, 2018). It avoids the problem of overfitting by taking the average of all the predictions, which cancels out the biases (DataCamp, 2016). However, it is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to predict the same given input and then perform voting on it. This whole process is time-consuming.

Yet, there are indeed a number of potential opportunities in burnt area mapping for ML applications where ML are yet to be applied or is underutilized. For instance, a method that could be used to improve weather station observations and for forecasting drought in the context of fire danger potential. It can also be useful in improving decision making (Jain et al., 2020). Smoke detection, which is important for fire detection, and determining the presence of false negatives in hotspot data could both benefit from a similar approach (ibid.).

CONCLUSION

In conclusion, though visually Random Forest and well-parameterized Support Vector Machines gave more or less similar results but numerical Random Forest was better. The burned areas were distinguished well in this analysis. However, features with spectral behaviour similar to that of the burned area caused by topography shadows and changes in land cover not related to fires were wrongly classified. Therefore, it was realized that special care should be taken in areas where these characteristics and events occur close to the fire-affected area. As a result, assessing the separability for different classes of land use and the influence of sample size as a future study in the study area could be a good alternative. In order to make my model more robust, I should consider an improved dataset and also gain additional knowledge in mapping fire. On the other hand, even though ML models can learn on their own, expertise in wildfire science is required to ensure realistic wildfire analysis, and the complexity of some ML methods necessarily requires a specialized and structure formed of their application (Jain et al., 2020).

Nevertheless, there is no such thing as the best machine learning algorithm. Each method is unique and is adapted to the available data, the context of the domain problem, and any external/internal constraints. If you're wondering, "So... which model should I use?" the solution is to test as many as you can and then analyze what works best for you.

REFERENCES

Bar, S., Parida, B. R., & Pandey, A. C. (2020). Landsat-8 and Sentinel-2 based Forest fire burn area mapping using machine learning algorithms on GEE cloud platform over Uttarakhand, Western Himalaya. *Remote Sensing Applications: Society and Environment*, *18*(May). https://doi.org/10.1016/j.rsase.2020.100324

Belgiu, M., & Drăgu, L. (2016, April 1). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 114, pp. 24–31. Elsevier B.V. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman, L. (2001). *Random Forests* (Vol. 45).

Camps-Valls, G. (2009). Machine learning in remote sensing dataprocessing. *Machine Learning for Signal Processing XIX - Proceedings of the 2009 IEEE Signal Processing Society Workshop, MLSP 2009*, (November). https://doi.org/10.1109/MLSP.2009.5306233

cloudml. (2018, February 22). Learn 7 Machine Learning Algorithms in 7 Minutes - machine learning data science guide tutorial process. Retrieved June 15, 2021, from https://www.claoudml.com/blog/machine-learning-made-easy

Datacamp. (2019). (Tutorial) Support Vector Machines (SVM) in Scikit-learn - DataCamp. Retrieved June 26, 2021, from https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

DataCamp. (2016, May 16). Random Forests Classifiers in Python - DataCamp. Retrieved June 26, 2021, from https://www.datacamp.com/community/tutorials/random-forests-classifier-python

Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). *A review of machine learning applications in wildfire science and management*. Retrieved from http://arxiv.org/abs/2003.00646

Koehrsen Will. (2018, January 10). Hyperparameter Tuning the Random Forest in Python | by Will Koehrsen | Towards Data Science. Retrieved June 26, 2021, from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *Https://Doi.Org/10.1080/01431160600746456*, *28*(5), 823–870. https://doi.org/10.1080/01431160600746456

Mahmoud, M. A. I., & Ren, H. (2019). Forest fire detection and identification using image processing and SVM. *Journal of Information Processing Systems*, *15*(1), 159–168. https://doi.org/10.3745/JIPS.01.0038

Pacheco, A. D. P., Junior, J. A. D. S., Ruiz-Armenteros, A. M., & Henriques, R. F. F. (2021). Assessment of k-nearest neighbor and random forest classifiers for mapping forest fire areas in central portugal using landsat-8, sentinel-2, and terra imagery. *Remote Sensing*, *13*(7), 1–25. https://doi.org/10.3390/rs13071345

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*(5), 1007–1011. https://doi.org/10.1080/01431160512331314083

Schepers, L., Haest, B., Veraverbeke, S., Spanhove, T., Borre, J. Vanden, & Goossens, R. (2014). Burned area detection and burn severity assessment of a heathland fire in belgium using airborne imaging spectroscopy (APEX). *Remote Sensing*, *6*(3), 1803–1826. https://doi.org/10.3390/rs6031803

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

Singh, A. (2020). *Case Study on 2019 Australian Bushfire*. Retrieved from https://www.researchgate.net/publication/340930739_Case_Study_on_2019_Australian_Bushfire

Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors (Basel, Switzerland)*, *18*(1). https://doi.org/10.3390/s18010018

Yıldırım Soner. (2020, May 31). Hyperparameter Tuning for Support Vector Machines — C and Gamma Parameters | by Soner Yıldırım | Towards Data Science. Retrieved June 27, 2021, from https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167