A Subspace KNN Ensemble Classifier for Land Cover Classification using Medium Resolution Satellite Images

Atijosan Abimbola
COPINE, National Space Research and Development Agency,
Obafemi Awolowo University Campus,
Ile-Ife, Osun State, Nigeria

Olaoluwa Idayat A.
COPINE, National Space Research and Development Agency,
Obafemi Awolowo University Campus,
Ile-Ife, Osun State, Nigeria

ABSTRACT
Accurate land cover information is crucial for the development and implementation of various environmental, socio-political and economic policies. Classification techniques are fundamental in this regard, as they determine the accuracy of information obtained from land cover classification and thus, affect the accuracy of subsequent applications.

In this research, a subspace KNN ensemble classifier with a nearest neighbour learning algorithm is proposed for the accurate classification of medium-resolution multispectral satellite images. The Landsat satellite dataset obtained from the UC Irvine machine learning repository was used as the testing data. For performance evaluation, confusion matrix and receiver operating curve plots were used for performance evaluation.

Performance comparison was made with three well-known machine learning classifiers namely, Decision Tree (DT), Support Vector Machine (SVM) and Kernel Naïve Bayes (KNB) models to determine the model with the highest accuracy. Results obtained show that the subspace KNN ensemble classifier outperforms the other classifiers in terms of accuracy as it achieves a 91.5% accuracy while DT, SVM and KNB classifiers achieved 85%, 90.4 and 81.8% accuracy respectively.

Keywords
Subspace KNN ensemble classifiers, Support Vector Machines, Medium Resolution Satellite Images, Land Cover Classification

1. INTRODUCTION
Accurate land cover information represented on thematic maps is crucial for various uses ranging from environmental conservation and management to socio-political and economic applications (Phan et al 2020; Talukdar et al 2020; Atijosan et al., 2016). Furthermore, land use information is of great relevance to policy formulation and implementation across all levels of governance and development (Tariq and Mumtaz, 2023, Alam et al 2020). The aim of satellite image classification for land cover uses is to accurately categorize all pixels in the satellite image into the distinct classes (e.g. forest, water, crops etc.) that they represent. Pixel categorization is carried out using classification techniques. Classification techniques are very crucial in this process as they determine the accuracy of information obtained from land cover classification and thus, affect the accuracy of all subsequent applications (Phan et al 2020). Therefore, the development of accurate classification techniques for land cover information extraction is important and in high demand (Phan et al 2020).

Remotely sensed satellite images are recognized as one of the most important data sources for land cover mapping and for monitoring land cover change over time (Phan et al 2020). Medium-resolution satellite imageries are the most important and widely used data source for producing maps of LULC over large areas due to their inherent ability to provide near-global coverage of the Earth's surface at a high frequency (Saini and Rawat, 2023, Ali and Johnson 2022). Landsat medium-resolution multispectral satellite images are widely used data sources for land cover mapping and land cover change monitoring (Ouchra et al 2023, Bouslihim et al 2022).

Recently, the application of machine-learning algorithms on remotely-sensed satellite imageries for land cover classification has been attracting considerable attention (Saini and Rawat, 2023, Ali and Johnson 2022). Machine learning algorithms offer the prospects for more effective and efficient classification of remotely sensed imagery (Ali and Johnson 2022; Maxwell et al 2018). Furthermore, studies have generally found that they tend to produce higher accuracy compared to traditional classifiers (Fotso Kamga et al 2021; Singh and Tyagi, 2021). Their strengths include the capacity to handle data of high dimensionality and to map classes with very complex characteristics (Maxwell et al 2018).

Machine learning classification methods based on training several heterogeneous models and then aggregating their predictions using certain strategies tend to provide more effective solutions to the classification problem (Herrera et al 2016). These types of classifiers are widely known as machine learning-based ensemble classifiers. They are known to help improve classification performance when compared with single classifiers (Kiziloz 2021). The main goal of an ensemble classifier is to minimize the misclassification rate of a weak classifier by combining multiple classifiers (Schneider and Xhafa 2022). The basic idea is to obtain predictions of multiple classifiers on the original data and combine the different predictions to make a strong classifier. Leveraging the power of subspace methods, this research work presents a subspace KNN ensemble classifier that will further enhance the accuracy of land cover information obtained from land cover classification using medium-resolution satellite images. The ensemble classifier incorporates the collective intelligence of multiple subspace KNN models, thus, fostering a robust and adaptable system for land cover classification. This approach not only capitalizes on the strengths of individual models but also provides a mechanism for handling diverse land cover patterns present in medium-resolution satellite imagery thus enhancing the accuracy and efficiency of the land cover classification. The performance of the subspace ensemble KNN model for land cover classification of medium resolution satellite images is compared with three well-known machine learning classifiers namely, Decision Tree Classifier, Support Vector Machine (SVM) classifier and Kernel Naïve Bayes classifier models to determine the effectiveness.

This paper contributes to this evolving intersection of machine learning and satellite image classification by introducing a subspace KNN ensemble classifier for medium-resolution satellite imagery that offers a promising avenue for more accurate land cover classification. The rest of the paper is organized as follows. Section 2 delves into the methodology and the proposed method. Section 3 reports on the results obtained and comparison with other ML classifiers. Finally, section 4 concludes the study

2. METHODOLOGY

The methodology used is detailed in this section.

2.1 Dataset

The dataset used was obtained from the UC Irvine machine learning repository (Srinivasan,1993). The Statlog Landsat satellite dataset consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a Landsat satellite image and the classification associated with the central pixel in each neighbourhood. In the database, the class of a pixel is coded as a number. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighbourhood and a number indicating the classification label of the central pixel. The number is a code for the following classes shown in Table 1.

Table 1: Dataset classes

Class Number	Class Description
1	Red soil
2	Cotton crop
3	Grey soil
4	Damp grey soil
5	Soil with vegetation stubble
6	Mixture class (all types present)
7	Very damp grey soil

NB. There are no examples with class 6 in this dataset.

Figure 1 shows the scatter plot of the dataset. The scatter plot shows the distribution of the six classes {class 1 (red soil), class 2 (cotton crop), class 3 (grey soil), class 4 (damp grey soil), class 5 (soil with vegetation stubble), class 7 (very damp grey soil)} present in the dataset.

2.2 Classification Model Validation Scheme

Validation helps us choose the best models and protects against overfitting. A Fivefold cross-validation scheme was used as the default validation scheme for all classifier models used in this study.

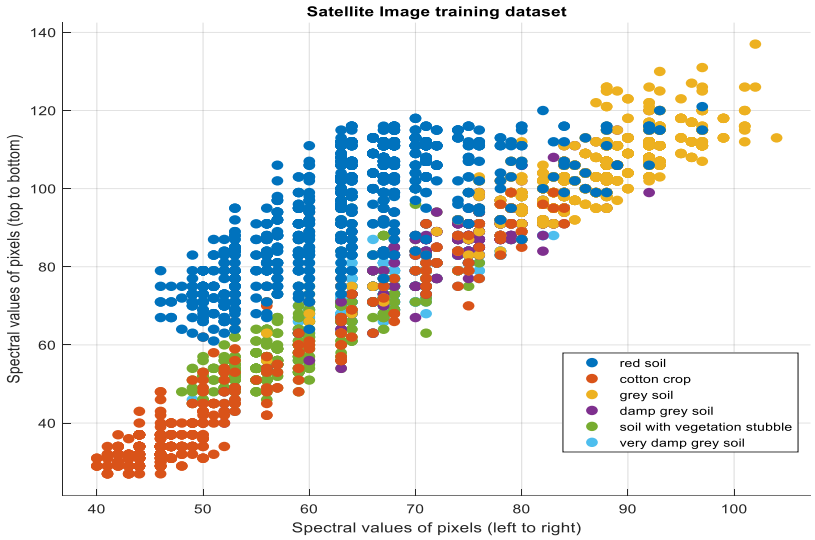

Figure 1. Scatter plot of the dataset

2.3 Ensemble classifier

A basic principle behind ensemble learning is the attempt to combine a series of classifier models produced by several learners into an ensemble that will perform better than the original learners (Jozdani et al., 2019). The ensemble classifier simply develops a strong classifier by integrating the output of several weak or base classifiers to enhance the overall classification accuracy (Jozdani et al 2019). The ensemble classifier can consist of any type of base classifier algorithm such as K-nearest neighbour (KNN) or other sorts of base learner classification algorithms. In this study, KNN and random subspace were utilized as the base learner and ensemble approach respectively.

The KNN algorithm is essentially a machine learning method that divides the feature space into distinct clusters based on the features associated with the different classes. This classifier considers the k metric distances between the test sample features and those of the nearest classes while classifying a test feature vector (Rashid et al 2019). While k-NN is stable in terms of adjusting training datasets, it is susceptible to feature set variation. Due to the perceptive nature of the input selection of KNN, ensemble systems based on random subspaces are capable of enhancing the efficiency of single KNN classifiers.

Random subspace is a frequently utilized ensemble technique that generates individual classifiers from randomly chosen subspaces of data (Rashid et al 2019). The learner type used in this study is Nearest Neighbour with 30 numbers of learners and 2 subspace dimensions.

2.4 Evaluation parameters

Evaluation of classification algorithms is one of the key points in any process of data mining. Two commonly used tools in analysing the results of classification algorithms are the confusion matrix and Receiver Operating Curves (ROC) plot (Oprea and Ti 2014).

2.4.1 Receiver Operating Curves (ROC) plot

The ROC curve graphically displayed the binary classification model's performance. To interpret the ROC curve, Area Under the Curve (AUC) values will be considered. An AUC of 1.0 suggests a perfect model fit (Liyanage et al 2023).

2.4.2 Confusion matrix

The confusion matrix provides a performance summary of the classification model by comparing predicted and actual values. The matrix comprised rows and columns, with each row representing instances in a predicted class

and each column representing instances in an actual class (Liyanage et al 2023). True Positive Rates (TPR) and False Positive Rates (FPR) will be considered.

2.5 Performance evaluation with different Machine Learning based (ML) classifiers

Three machine learning-based classifiers were used to classify the dataset and their performance where compared with the subspace KNN-based ensemble classifier. The classifiers are namely;

2.5.1 Decision Tree Classifier (DTC)

Decision tree classifiers have been widely used for land cover classification (Purwanto et al 2023). The DTC is a machine-learning-based classifying technique that comprises several classes of modelling algorithms using a tree-like structure, where each node shows a test on attributes, branches represent the test results, and the leaf node shows the target classes (Purwanto et al 2023).

2.5.2 Support Vector Machine

SVM is one of the most commonly used classifiers in the ML community that categorizes data using an optimally separating hyperplane. One key advantage of SVM for remote sensing applications is its inherent ability to handle high-dimensional data using relatively few training samples (Jozdani et al., 2019).

2.5.3 Naïve Bayes Classifier

A naive Bayes classifier is a probabilistic classifier which uses Bayes' theorem to assign events to classes. The classification of an unknown event is made by comparing the attribute values of the events with the statistics of each class. The class with the highest similarity is chosen. This classifier has proven to work well on a variety of problems and it is considered a very effective supervised classifier due to its high level of accuracy and low computation time (Sa'idah et al., 2019; Cervone and Haack 2012)

3. RESULTS AND DISCUSSION

Results obtained and discussion are presented in this section.

3.1 Subspace KNN classification parameters

Classification learner in MATLAB machine learning and deep learning app was used to implement the classifiers and carry out the classification of the dataset. The subspace KNN classifier parameters are highlighted in Table 2. Table 2 consists of the preset, ensemble method, learner type, number of learners and subspace dimension used.

Table 2: Subspace KNN classification parameters

Preset	Ensemble method	Learner type	No of learners	Subspace dimension
Subspace KNN	Subspace	Nearest Neighbours	30	30

Table 3 highlights the results obtained. Classification accuracy was 91.3 %. The total misclassification cost was 384 and the prediction speed and training time were ~1100 obs/sec and 18.654 seconds respectively. Scatter plots are important has they help to show the relationship between two variables. The scatter plot of the classified class data is shown in Figure 2. Figure 2 shows both the correct and incorrect model prediction for all the variables.

Table 3: Results obtained from the classification using subspace KNN classifier

Accuracy	Total misclassification cost	Prediction speed	Training time
91.3 %	384	~1100 obs/sec	18.654 sec

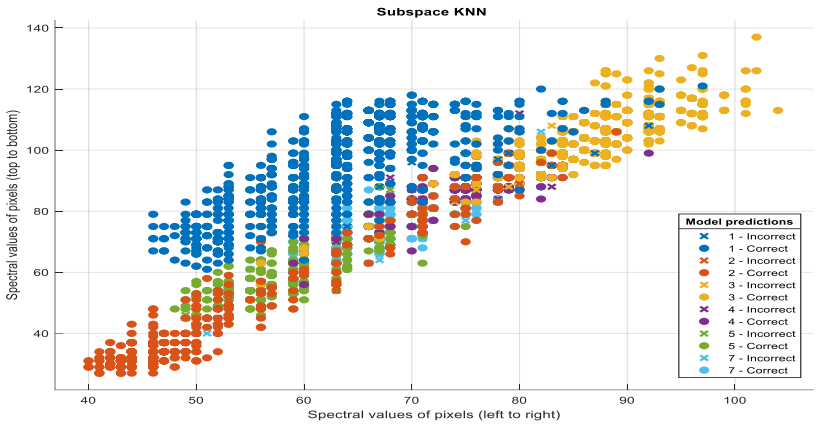

Figure 2: Scatter plot after classification

3.2. Comparison with different classifiers

Results obtained from the performance comparison of the different classifiers are presented in Table 4. Classification accuracy refers to the percentage of observations that are correctly classified. Higher values point to more accurate classification. From the results shown in Table 4, the Subspace KNN ensemble had a classification accuracy of 91.5% which is the highest compared with the accuracy of the other three classifiers namely, Decision Tree (85%), SVM (90.4%) and Naïve Bayes (81.8%). From Table 4 it can also be observed that the misclassification cost for the Subspace KNN ensemble classifier was the lowest at 384 when compared with the other three classifiers, Decision Tree (665), SVM (427) and Naïve Bayes (807).

AUC values correspond to the integral of a ROC curve TPR values with respect to FPR, from FPR=0 to FPR=1. The AUC is a measure of the quality of a classifier. Comparison of the AUC values for the four classifiers across the six classes described in Table 1 are shown in Table 4. Subspace KNN ensemble and SVM classifiers AUC values across all six classes were the highest compared with the other two classifiers (SVM and Naïve Bayes). An AUC of 1.0 suggests a perfect model fit. Thus, the closer the AUC value is to 1, the better the classifier. The ROC plots over class 5 in the Landsat dataset for Subspace KNN ensemble, Decision Tree, SVM and Naïve Bayes classifiers are shown in Figures 3,4,5 and 6 respectively.

Table 4: Performance comparison

Classifier	**Subspace KNN ensemble**	**Decision Tree**	**SVM**	**Naïve Bayes**
Model type	Subspace KNN	Fine Tree	Quadratic SVM	Kernel Naïve Bayes
Accuracy	91.5 %	85%	90.4	81.8%
Misclassification cost	384	665	427	807
AUC for class 1 (Red soil)	1	0.97	1	0.97
AUC fro class 2 (Cotton crop)	0.99	0.98	0.99	0.99
AUC for class 3 (Grey soil)	0.99	0.96	0.99	0.97
AUC for class 4 (Damp Grey soil)	0.95	0.85	0.95	0.87
AUC for class 5 (Soil with vegetation stubble)	0.98	0.90	0.98	0.93
AUC for class 7 (Very damp grey soil)	0.99	0.95	0.98	0.92

Confusion matrix plots of the classifiers (Subspace KNN ensemble, Decision Tree, SVM and Naïve Bayes) are shown in figures 7, 8, 9 and 10 respectively. Summary of the classifiers performance per class are shown in the last two columns on the right of each plot (figures 7, 8, 9 and 10). TPR refers to the proportion of correctly classified observations per true class while FNR refers to the proportion of incorrectly classified observations per true class. On average across the six classes, the percentage TPR value of the Subspace KNN ensemble classifier exceeds that of the other classifiers.

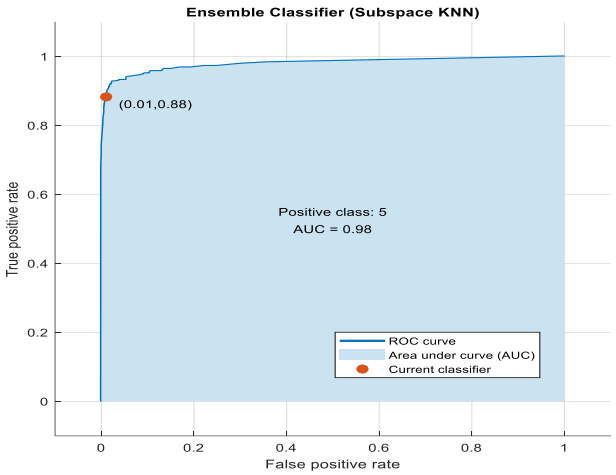

Figure 3: Receiver Operating Curves (ROC) plot of Ensemble Subspace KNN Classifier

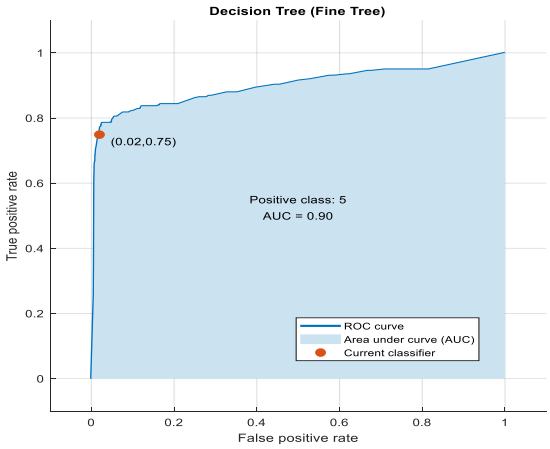

Figure 4: Receiver Operating Curves (ROC) plot of Decision Tree Classifier

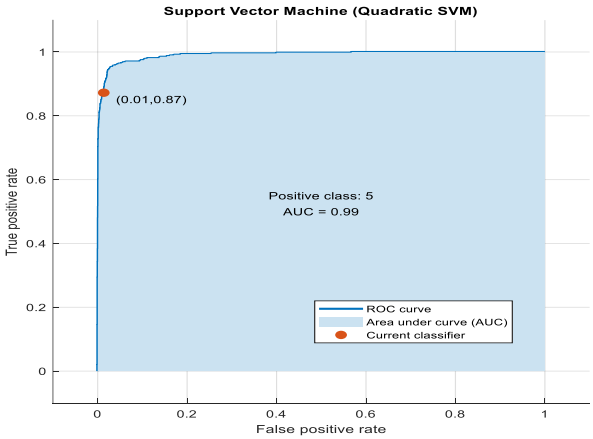

Figure 5: Receiver Operating Curves (ROC) plot of SVM Classifier

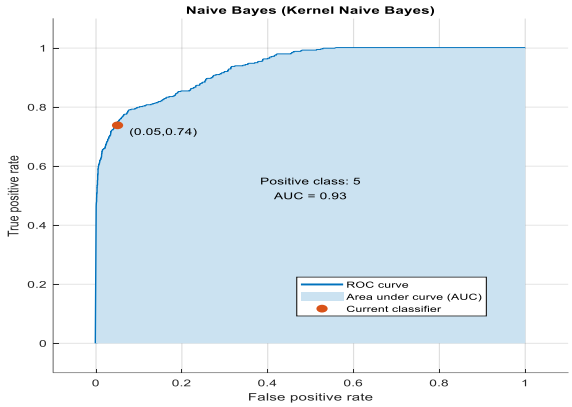

Figure 6: Receiver Operating Curves (ROC) plot of Naïve Bayes Classifier

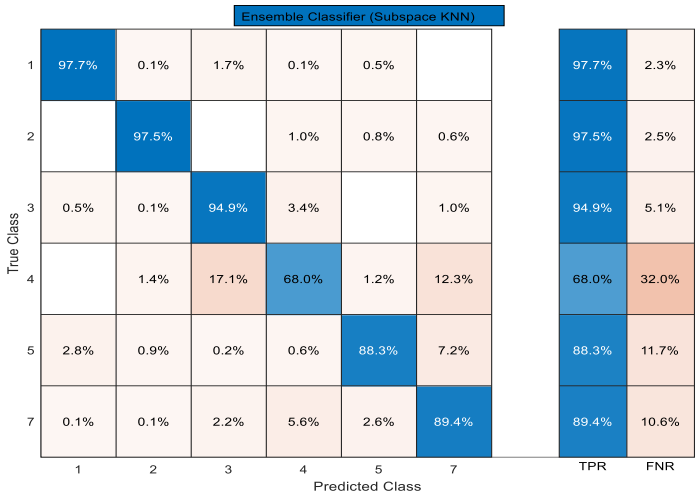

Figure 7: Confusion matrix of Ensemble KNN Classifier (True positive rates and False negative rates)

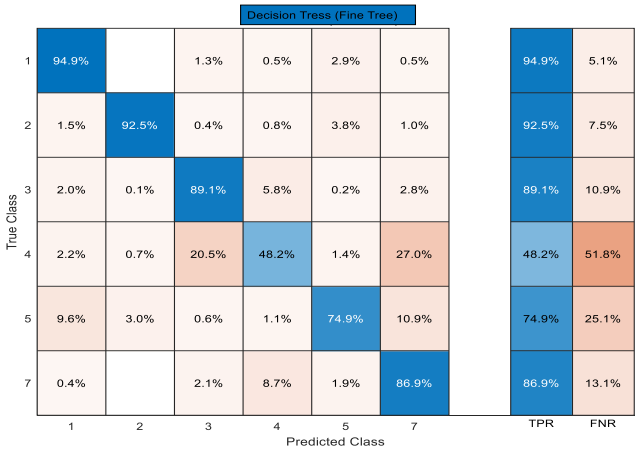

Figure 8: Confusion matrix of Decision Tree Classifier (True positive rates and False negative rates)

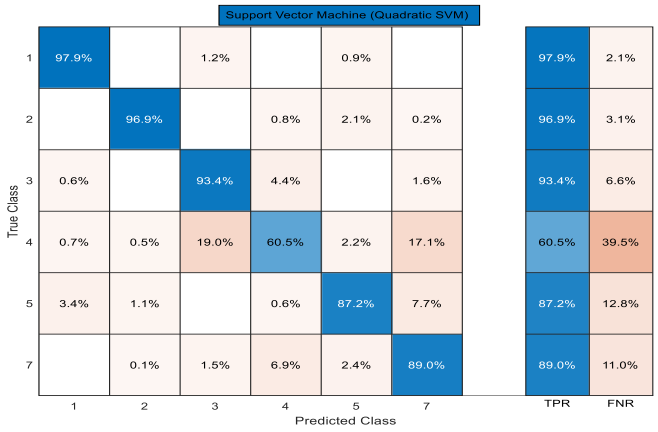

Figure 9: Confusion matrix of Support Vector Machine Classifier (True positive rates and False negative rates)

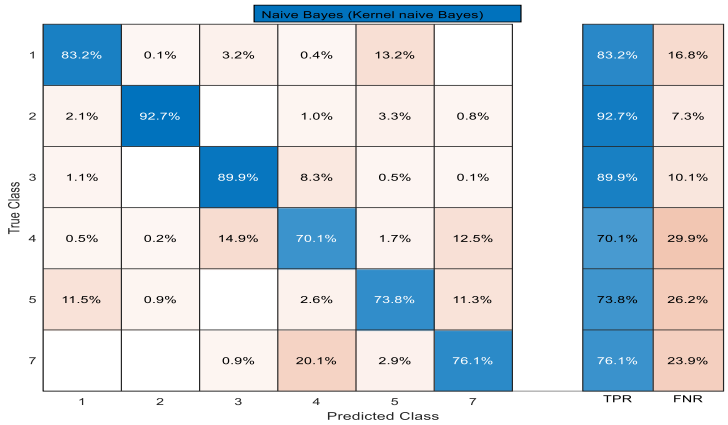

Figure 10: Confusion matrix of Naïve Bayes Classifier (True positive rates and False negative rates)

4. CONCLUSION

Satellite imageries are an important data source for land cover classification. In this study, a subspace KNN ensemble classifier was developed and used to classify the Landsat dataset obtained from a publicly available machine learning repository. Confusion matrix and ROC plots were used for performance evaluation. Performance comparison was also carried out by comparing the classification accuracy of the subspace KNN ensemble classifier with that of Decision Tree, SVM and Kernel Naïve Bayes classifiers. Results obtained showed that the accuracy of the subspace KNN ensemble classifier outperforms the other three classifiers. Further research would be directed to applying this technique to the classification of high-resolution satellite images.

5. REFERENCES

[1] Alam, A., Bhat, M. S., & Maheen, M. (2020). Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley. GeoJournal, 85, 1529-1543.

[2] Ali, K., & Johnson, B. A. (2022). Land-Use and Land-Cover Classification in Semi-Arid Areas from Medium-Resolution Remote-Sensing Imagery: A Deep Learning Approach. Sensors, 22(22), 8750.

[3] Atijosan, A., Badru, R., Babalogbon, A., & Alaga, T. (2016). Classification of Medium Resolution Satellite Imageries using Artificial Neural Network and Swarm Intelligence. International Journal of Hybrid Information Technology, 9(11), 215-228.

[4] Bouslihim, Y., Kharrou, M. H., Miftah, A., Attou, T., Bouchaou, L., & Chehbouni, A. (2022). Comparing pan-sharpened landsat-9 and sentinel-2 for land-use classification using machine learning classifiers. Journal of Geovisualization and Spatial Analysis, 6(2), 35.

[5] Cervone, G., & Haack, B. (2012). Supervised machine learning of fused RADAR and optical data for land cover classification. Journal of Applied Remote Sensing, 6(1), 063597-063597.

[6] Fotso Kamga, G. A., Bitjoka, L., Akram, T., Mengue Mbom, A., Rameez Naqvi, S., & Bouroubi, Y. (2021). Advancements in satellite image classification: methodologies, techniques, approaches and applications. International Journal of Remote Sensing, 42(20), 7662-7722.

[7] Herrera, F., Charte, F., Rivera, A. J., del Jesus, M. J., Herrera, F., Charte, F., ... & del Jesus, M. J. (2016). Ensemble-based classifiers. Multilabel classification: problem analysis, metrics and techniques, 101-113.

[8] Jozdani, S. E., Johnson, B. A., & Chen, D. (2019). Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification. Remote Sensing, 11(14), 1713.

[9] Kiziloz, H. E. (2021). Classifier ensemble methods in feature selection. Neurocomputing, 419, 97-107.

[10] Liyanage, V., Tao, M., Park, J. S., Wang, K. N., & Azimi, S. (2023). Malignant and non-malignant oral lesions classification and diagnosis with deep neural networks. Journal of Dentistry, 137, 104657.

[11] Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International journal of remote sensing, 39(9), 2784-2817.

[12] Oprea, C., & Ti, Ș. (2014). Performance evaluation of the data mining classification methods. Information society and sustainable development, 1, 249-253.

[13] Ouchra, H. A. F. S. A., Belangour, A., & Erraissi, A. L. L. A. E. (2023). Machine learning algorithms for satellite image classification using Google Earth Engine and Landsat satellite data: Morocco case study. IEEE Access.

[14] Phan, T. N., Kuch, V., & Lehnert, L. W. (2020). Land cover classification using Google Earth Engine and random forest classifier. The role of image composition. Remote Sensing, 12(15), 2411.

[15] Purwanto, A. D., Wikantika, K., Deliar, A., & Darmawan, S. (2023). Decision tree and random forest classification algorithms for mangrove forest mapping in Sembilang National Park, Indonesia. Remote Sensing, 15(1), 16.

[16] Rashid, M., Mustafa, M., Sulaiman, N., Abdullah, N. R. H., & Samad, R. (2021). Random Subspace K-NN Based Ensemble Classifier for Driver Fatigue Detection Utilizing Selected EEG Channels. Traitement du Signal, 38(5).

[17] Sa'idah, S., Pratiwi, N. K. C., Aprilia, B. S., Magdalena, R., & Fu'adah, Y. N. (2019, November). Land cover classification using Grey Level Co-occurrence Matrix and Naive Bayes. In Journal of Physics: Conference Series (Vol. 1367, No. 1, p. 012073). IOP Publishing.

[18] Saini, R., & Rawat, S. (2023, March). Land Use Land Cover Classification in Remote Sensing Using Machine Learning Techniques. In 2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP) (pp. 99-104). IEEE.

[19] Schneider, P., & Xhafa, F. (2022). Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to EHealth and Patient Data Monitoring. Academic Press.

[20] Singh, M., & Tyagi, K. D. (2021). Pixel based classification for Landsat 8 OLI multispectral satellite images using deep learning neural network. Remote Sensing Applications: Society and Environment, 24, 100645.

[21] Srinivasan, Ashwin. (1993). Statlog (Landsat Satellite). UCI Machine Learning Repository. https://doi.org/10.24432/C55887.

[22] Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations. A review. Remote Sensing, 12(7), 1135.

[23] Tariq, A., & Mumtaz, F. (2023). Modeling spatio-temporal assessment of land use land cover of Lahore and its impact on land surface temperature using multi-spectral remote sensing data. Environmental Science and Pollution Research, 30(9), 23908-23924.