Machine Learning Algorithms for Extraction of Glacial Lakes Using Ground Range Detected (GRD) Data: A Case Study from Hunza River Basin, Pakistan

Hajra Nazakat (hajirahasni1998@gmail.com)
Karakoram International University https://orcid.org/0000-0003-2257-0760
Syed Najam ul Hassan
Karakoram International University
Garee Khan
Karakoram International University
Aftab ahmad
Karakoram International University
javed Akhter Qureshi
Karakoram International University
Sajid Ali
Institute of Tibetan Plateau Research Chinese Academy of Sciences

Research Article

Keywords: Batura, Classifiers, Glacial Lakes, Hunza river basin, Sentinel-1 GRD

Posted Date: August 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-590990/v1

Abstract

In the Karakoram Mountain range, glacial lakes are essential elements of the cryosphere. As a function of climate change and increasing temperature, these glacial lakes threaten downstream existence and the ecosystem by short time glacial lake outburst floods (GLOF). Therefore, the Glacial Lake mapping technique is a vital task to observe GLOF hazards. In this study, microwave Sentinel-1 Ground Range Detected (GRD) data used. It has the dual-polarization capability (HH + HV or VV + VH) and the ability to penetrate even through clouds or any weather condition. The study objective is to explore the application of GRD data and evaluate the efficiency and accuracy of machine learning algorithms for the extraction of water bodies. The study method is based on two main procedures, GRD backscattering analysis and supervised Machine Learning classifiers. The most commonly used machine learning classifiers are Random Forest (RF), K-nearest neighbor (KNN), and Maximum Likelihood. Although both procedures show better results for glacial lakes mapping in the study area, the mean backscatter parameter has the best accuracy rate than others in the total backscattering analysis. Likewise, in the classification approach, accuracy assessment was executed by comparing the results obtained for each classifier with the reference data. For all experiments, KNN performed the best at given training samples (Accuracy = 93%, Error rate = 0.06%) for both classes, compared to RF (Accuracy = 92%, Error rate = 0.07) and Maximum Likelihood (Accuracy = 90%, Error rate = 0.09%). The high classification accuracy obtained to extract glacial lakes using our approach will be useful to determine the short time flood outburst and take future precautionary measurements.

Introduction

The extraction of a glacial lake is the process of monitoring, mapping, and detecting water bodies from glaciated mountains using satellite images(Mitkari, Arora, & Tiwari, 2017). This process of glacial lakes monitoring is a crucial task that plays an important role in reducing natural disasters, global warming, and various human threats on a regional and global scale (Verpoorter, Kutser, & Tranvik, 2012). In the last three decades, glacial lakes are rapidly growing worldwide due to climate change and glacier retreats (Glacial Lakes Have Grown Rapidly Worldwide- Satellite Images _ Earth, n.d.). According to Shugar et al. (Shugar et al., 2020), the glacier lakes volume increased by around 48% to 156.5km3 between 1990 and 2018 globally. So, the increasing glacial lakes are not only the most vital climate indicators and water resources, but they also act as a cause of many glacial hazards such as Glacial Lake Outburst Floods GLOFs (Yao, Liu, Han, Sun, & Zhao, 2018).

Similarly, at a regional scale, Hindukush Karakorum and Himalayan (HKH) mountains are mainly sensitive to climate change therefore, the number of glaciers and glacial lakes is constantly changing with time and expanding or born new glacial lakes rapidly (Ashraf, Naz, & Iqbal, 2017). However, in the late 1990s, the big debate "Karakorum Anomaly" arises among scientific societies and researchers that many glaciers of the central Karakoram region are stable or advances which is peculiar behavior than rest of world glaciers(Hewitt, 2005). Later, (Bazai, Cui, Carling, & Wang, 2020)(Ashraf et al., 2017) found that glacier and glacial lakes in various river basins indicated different patterns (may stable or may not)

depending on geographic location in the HKH region. Like in the Hunza river basin of Karakorum range, there were five GLOF events that occurred during 2007 and 2008, which harshly affected the nearby societies and modeled a threat for the future(Ashraf, Naz, & Roohi, 2012). Thus, the monitoring of glacial lakes on a regional and local scale is therefore an enormous function. Because it plays a significant role in natural hazard prevention, global warming, and varied human therapies (Verpoorter et al., 2012). Other than that, glacier lake mapping will mitigate risks to GLOFs that are very risky to downstream including infrastructure and vegetation. (Yasmeen, 2013).

All the above-mentioned (Mitkari et al., 2017)(Yasmeen, 2013) studies indicate that the significance of mapping glacial lakes from high altitude mountain is more important and need of time. Accordingly, the research (Elsahabi, Negm, & Ali, 2016) was performed to extract glacial lakes from steeped mountains using multiple forms of data and techniques. Such as the study(Chen, Zhang, Tian, & Li, 2017) performed to extract glacial lakes annually in the entire Tibet Plateau region and were used Landsat-8 images and non-local active contour techniques. The authors(Chen et al., 2017) noticed that the method introduced is theoretically applicable to large-scale initiatives related to the mapping of water bodies. On the other side, Landsat 8 has a cumulative research area of 35.2 observations. Out of overall measurements, 21.2 measurements are decent and reliable regardless of cloud cover. Likewise,(Wangchuk & Bolch, 2020) proposed mapping glacial lakes and frozen glacial surfaces. Their method used comparatively multi-source data such as SAR and optical. They found that different factors challenge the mapping of water bodies in the alpine region. These factors include the shadow of mountains, cloud cover in optical data, and small glacial lakes. Thus, most of the studies used visual remote sensing data (Landsat, sentinel-2A/B, Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), and Moderate Resolution Imaging Spectroradiometer (MODIS), etc.) to monitor the spatial and temporal changes of glacial lakes (Zhang, Chen, Tian, Liang, & Yang, 2019a). This optical data is only adequate during the absence of clouds. However, clouds are mainly appearing in high-altitude mountains, especially in the Karakoram Mountains of northern areas. In this case, the use of optical data is more challenging as some information related to an area of interest might be missing just because of cloud cover. Hence, to address optical data, we need such a type of Satellite data that penetrates even through clouds or any kind of weather condition in mountainous areas. At the same time, that is "Microwave sentinel-1 SAR GRD" data. Moreover, to avoid cloud cover in optical data, we have to take more additional processing steps. This type of processing may be time-consuming processing (Wangchuk & Bolch, 2020). Therefore, our study uses or contributes Sentinal-1 GRD data in an alternative way as it is independent of all-weather conditions. Apart from that, our study also helps to reduce the negative impact on livelihood by giving future precautionary measurements of GLOFs events.

Additionally, machine learning algorithms mainly offer an effective and efficient classification of remote sensing data. It can handle high-dimensional datasets and map classes with complex characteristics (Maxwell, Warner, & Fang, 2018). Therefore, our study comparatively analyzes some specific machine learning algorithms that can enhance the performance analysis of GRD images in more reliable way. The main objectives of this study are: 1) aimed to explore the importance and application of GRD data for extraction of glacial lakes. 2) to modify a conventional method for effective and automated analysis of

SAR imageries. 3) to evaluate the efficiency and accuracy of machine learning algorithms for the extraction of water bodies in mountainous areas.

Study Area

Mainly, the north part of Pakistan is composed of mountainous regions. The high mountain chain of Hindukush, Karakoram, and Himalayas (HKH) is home to several glaciers and glacial lakes in Gilgit-Baltistan. Accordingly, the GB Hunza basin is located in the extreme northern part of the Upper Indus Basin (UIB) at the junction of four high mountains. These mountains include Pamir, Tianshan, Karakoram, and Hindukush near the Pakistan-China border. The Hunza basin lies between latitude and longitude of 35^0 N to 36^0 N and 75^0 E to 76^0 E, respectively. The total area of the basin is 13,567.23 $km^{2,}$ as calculated in ArcMap.

Additionally, the Hunza basin glaciated area lies between 2280m to 7850m (*Second Summer School on Integrated Water Resources Management 27–31*, 2018). The Batura Glacier is an area of interest selected from the Gojal region of the Hunza basin. It is the fifth-longest glacier outside the polar regions as 57km its length, and spread over 285sqkm.

Data And Methods

Data Used

In this study, the essential step was the collection and preparation of data. Various type of data was collected, processed, and analyzed using different techniques and tools. These various types of data include Sentinel-1A (Synthetic Aperture Radar (SAR)) level-1 Ground Range Detected (GRD) data of 2019-09-29, Shapefile of Boundary as Vector data, and Digital Elevation Model (DEM). SAR data was available through Google Earth Engine (GEE), which the United States Geological Survey supports. Sentinel-1 GRD time series product was downloaded and analyzed to identify glacial lakes in the study area. The acquired data had interferometric wide swath (IW) sensor mode. Its polarization capability is dual (HH + HV or VV + VH), which can provide maximum ground surface information. The revisit-cycle of collected data is 12-days, and its resolution is 10m.

Furthermore, an essential thing about sentinel-1A data is that it can afford any weather condition while using. Additionally, the boundary of the Study area was available from Pakistan's topographic Sheets. Table 1. illustrates the description of these data sources.

Table 1
Data Source and Description

Data Type	Source	Resolution	Usage Description	Revisit Time	Acquisition Date	Pass Direction
SAR Level-1 (GRD)	Google Earth Engine website (supported by USGS)	10×10m	To identify glacial lakes	12-days	2019-09-29	Ascending
DEM	ASTER-GDEM of the USGS	30×30m	To define the topography of the area	-	-	-

Method

The overall workflow of the proposed method study is presented in Fig. 2. The software tools used for this study include Google Earth Engine (GEE), ArcGIS, ESA SNAP, and Google Earth Pro. GEE is used for the acquisition of GRD data with the desired clip. The ArcGIS 10.5.1 tool is also used for making the study area map with the help of vector files. Moreover, the ESA SNAP 7.0 is used for data pre-processing and post-processing. Glacial lakes and glacial surface correlation Graphs were designed in Origin-Pro. Manual digitization of lakes for validation was done in the Google Earth Pro tool.

The foremost step in the study was to clip the desired data or Area of Interest (AOI) located in the Hunza Basin boundary. This step was easily done in Google Earth Engine (GEE) by a python script.

Sentinel-1 GRD Data Pre-processing

SNAP (Sentinel Application Platform) is mainly used for SAR images pre-processing as well as post-processing. SAR data was accessible with a zip file (Liu, 2016).

Before applying terrain correction, the initial step 'Radiometric calibration was done to get radar backscattering coefficient sigma naught ($\sigma 0$) data by converting digital numbers. It corrects errors caused by the azimuth, height, and atmospheric conditions of the sun. Furthermore, with the help of pixel arrangements, it can easily describe an image's real content and spatial structure. Afterward, a Geometric/Terrain Correction was performed. This step generates a pretty much stretched image by proper map projection or coordinate system. The advantage of this step was that it overcomes the map projection problem.

The second important step was View Angle Correction. In this step, the angle of the desired sentinel image is selected as a view angle of collected sentinel satellite data was ranging from 29^0 to 46^0.

A good backscatter analysis and interpretation of GRD data need a filter called speckle filtering. So, Lee-sigma, speckle filter used in our work to remove inconsistent and invaluable data such as mountain or a cloud's shadow. Besides, the purpose of the Lee-sigma filter was to eliminate the speckle noise and show more clear lakes boundary as its efficiency is suitable than other filters.

Backscatter Analysis

In our study methodology, SAR backscatter analysis got two backscatter polarizations during calibration of GRD data, such as VH (vertical diffuse and horizontal receive) and VV (vertical diffuse and vertical receive). Those polarizations were further analyzed by applying some specific backscatter estimators. The determined estimators in SNAP include Mean, standard deviation, maximum, minimum (for VH and VV), maximum-minimum ratio, and polarized ratio. Some equations of these estimators are given below:

Standard deviation

Standard deviation was calculated by using Eq. (1).

$$stdev = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (10 * \log_{10}\sigma_i)^2 - \left(\frac{1}{N} * \sum_{i=1}^{N} 10 * \log_{10}\sigma_i\right)^2}$$

1

Where N is the total number of images and σi is the coefficient backscatter image i.

Maximum-Minimum ratio

Complete implementation of maximum (VH and VV) and minimum (VH and VV) the ratio of both estimators are considered.

$$\max - \min \text{ ratio} = 10 * \log_{10}\left(\frac{\sigma^0 \text{Max}}{\sigma^0 \text{Min}}\right)$$

2

Where σ^0 max is Sigma0 Maximum band and σ^0 min is Sigma0 Minimum band.

Polarized ratio

The final estimator was a polarized proportion that was calculated by Eq. (3).

$$\sigma^0\mathrm{VHrVV} = \left(\frac{\sigma^0\mathrm{VH}}{\sigma^0\mathrm{VV}}\right)$$

3

Machine Learning Classifiers

This past decade, there has been growing interest in machine learning (ML). Cheaper computing power and low cost of memory speed up this tremendous interest (Omondiagbe, Veeramani, & Sidhu, 2019). Therefore, many tools and techniques are used to implement machine learning algorithms on different types of datasets. Likewise, SNAP is one of the best satellite data processing plat-form that play a vital role in applying machine learning classifiers. The present study focuses on three supervised ML classifiers available in SNAP to check their performance in discovering water bodies in the area of interest.

1. Random Forest Classifier

Random Forest (RF) algorithm is made up of a collection of classifiers that are based on trees. Therefore, it is called a decision tree ensemble classifier(Onojeghuo et al., 2018). It is a supervised machine learning algorithm that can be used for the task of regression and classification. In this study, to train the RF model, the system randomly selected 7000 training samples. The randomly selected samples are called In-Bag (IB) samples, while the remaining samples are called Out-Of-Bag Samples (OOB) used for cross-validation to evaluate how well the model worked.

Additionally, the random choice of training samples helps decrease the training dataset's overfitting (Khan et al., 2020). The system used some decision trees concerning random bootstrap datasets to build an RF classifier in this type of classification. Each tree produces a prediction, but the classifier selects the best prediction result from them with the help of majority voting and generates the final strong output.

2. K-Nearest Neighbors Classifier

KNN is just a relatively simple supervised machine learning algorithm used mainly for the classification process. It defines the data point according to the classification of its neighbor. In KNN," K" is a parameter that refers to the number of neighbors nearest to be included in the majority voting process. In this classification, parameter tuning chooses the correct value of" K" as it is necessary for better classification accuracy. Mostly, to avoid misunderstanding between different classes of data, an odd value of "K" is selected(Gomathi, Geetha Priya, & Krishnaveni, 2018).

Additionally, KNN is a successful technique used when classifying images. In this study, the system implemented KNN for image classification into two main classes. This classification is determined by the

distance between pixels of the training and "K" values. It follows the function of Euclidean distance given by Eq. 1

$$\text{Euclidan\_dist.} = \sqrt{\sum_{i=1}^{K} (Xi - Yi)^2}$$

4

Where (Xi-Yi) difference between two variables taken, summed, and squared for the pixel number "K."

3. Maximum Likelihood Classifier

Maximum Likelihood classification is another essential machine learning statistical method that is based on the Bayes theorem, and it can be used for image classification(Gomathi et al., 2018). During implementing this method in SNAP the training dataset gives a threshold value for both classes (lakes and non-lakes). M-LH calculates the likelihood probability of pixels on the base of that threshold value. Each pixel is categorized according to a given class based on the maximum probability of the class assigned.

Results

The behavior of glacial surface and lakes in backscatter polarization

The process of backscattering polarization provides two main backscatter polarized bands (Sigma0-VH and Sigma0-VV) during radiometric calibration of the GRD data. Subsequently, speckle noise reduction Lee Sigma (3×3) was processed that generate the result with different intensity values for both polarized bands. The intensity values of glacier lakes vary from 0.05 to 0.15 and 0.2 to 0.6 for VH and VV, respectively shown in "Fig .4". Thus, with the help of the filter process, both polarized bands sigma0-VH and sigma0-VV give clear identification about the glacial lakes of Batura shown in "Fig .3" with red circles. The red circles indicate water bodies that appear on the glacier surface.

The behavior of glacial lakes and surfaces is shown in "Fig. 4". The backscatter value for lakes is high, while the backscatter value for surfaces is very low. On the other hand, both polarized bands indicate some lakes below the glacier surface. Thus, sigma0-VH and sigma 0-VV can identify water bodies below the surface, not only the upper side of the glacier surface.

Detection of glacial lakes by backscatter estimators

To extraction of glacial lakes from the study area, the backscatter estimators were calculated at each pixel (Liu, 2016). In the line of analysis, the main estimators include Mean, standard deviation, maximum,

minimum, maximum-minimum ratio, and polarized ratio. The figure shows the result of each parameter with RGB color composite images, and their further details are given below:

1) Mean backscatter

VH and VV-Mean3 intensity were determined along with frequency in mean backscatter analysis. The window size used for mean and other parameters was 3x3.

The statistical consequence of Mean for both polarized bands VH and VV can be observed from "Fig. 5 (a, b)". It indicates that the intensity of mean is very high for the Batura glacier's water bodies compared to its glaciated surface because it can detect lakes above and below the glacier surface presented in graphs (a, b) of Fig. 6.

2) Standard deviation backscatter

Standard deviation was another essential factor in backscattering analysis that can be considered a statistical parameter. According to "Fig. 5(c, d)", the standard deviation of VH and VV has less detection performance for water bodies as its backscattering values are relatively the same for both lakes and surface. In other words, the constraint of this parameter was that it could not detect glacial lakes that exist below the surface of glaciers, unlike the mean estimator that is shown in "Fig. 6(c, d)".

3) maximum-minimum ratio

For further backscatter analysis of the study area image, maximum and minimum non-linear filters were applied. The performance and map of these filters are shown in "Fig. 5,6" Where it is labeled that the backscatter rate of distinct objects (Minimum3 and Maximum3) become high for lakes and low in the context of glacier's shallow. Moreover, some liquid bodies appear below the shallow, which these parameters can also identify. According to graphs (i, j) of "Fig. 6", the maximum-minimum ratio cannot considerably differentiate lakes and glacial areas because it has similar backscattering values for glaciers and water bodies. In contrast to the max-min ratio and individual minimum/maximum parameters, it is stated that separate parameters are best enough to extract water bodies.

4) Polarize ratio

Subsequently, the execution of the PR ratio shows a small number of water frames on the study area in "Fig. 5 (k)". So, it has a little bit higher backscattering values than the glaciated surface. Thus, lakes and surface backscattering values are approximately the same as observed in "Fig. 6 (k)".

A comparison of the above backscattering parameters specifies that Mean, Maximum, and Minimum non-linear filters give better results in glacial water bodies and surface classification. These filters have the aptitude to detect lakes below and above the glacial surface, unlike other filters. Moreover, they can remove part of the shadow and dark surface noises with the help of spectral and backscatter variability.

A. The experimental results of machine learning classifiers

The supervised classification algorithms required training samples from all possible classes. Therefore, the training samples were collected for two selected classes: lakes and glacial surfaces. In total, we collected 10,000 pieces, of which 5000 from a lake class and the remaining 5000 from a glacial surface class. The collected dataset was further split into two sets: training (70%) and testing (30%). So, 7000 samples were used for training, and 3000 models were used for testing, respectively.

In this study, to train the Random Forest model, the system selected 7000 training samples from which 4000 from the lake class and 3000 from the glacier surface class. These two classes are considered as training vectors in SNAP. According to the working mechanism of the RF classifier, our system gives the result by taking 116sec of processing time. Then, 92% classification accuracy was obtained for two different classes of the selected area. Similarly, the K-nearest neighbor classifier was implemented based on Euclidean distance for two main classes. It gives 93% classification accuracy for both distributions (lakes and glacier surface) by taking 50min processing time. Likewise, the Maximum Likelihood Classifier was executed, which generates 90% classification performance to extract glacier lakes from high altitude mountains glaciers. It took 38secconds for the process to complete.

Overall, three specific classifiers (RF, KNN, Maximum likelihood) were used in the current study. Since RF can easily handle outliers by combining trees as compared to the decision tree. Additionally, KNN is very simple and easy to implement. Moreover, it does not require a number of training before making predictions. At last, the M-LH classifier was used due to its scalability. It is highly scalable with several predictors and data points. The accuracy result of each classifier is summarized in Table 2.

Table 2
Classification results of classifiers

S. N0	Classifiers	Accuracy	Precision	Error rate	Recall
1	FR	0) 92%	0) 95%	0) 0.07%	0) 92%
		1) 92%	1) 89%	1) 0.07%	1) 93%
2	K-NN	0) 93%	0) 97%	0) 0.06%	0) 92%
		1) 93%	1) 88%	1) 0.06%	1) 96%
3	M-LH	0) 90%	0) 96%	0) 0.09%	0) 86%
		1) 90%	1) 82%	1) 0.09%	1) 95%

Table 3. Accuracy assessment of three selected classifiers for classification of lakes and glacial surface.

(0) represents lake class while (1) represents glacier surface

S.NO	Classifier	Accuracy (%)	Completion time (seconds)
1.	Fandom Forest	92%	116 sec
2.	K-Nearest Neighbor	93%	3000 sec
3.	Maximum Likelihood	90%	38 sec

B. Accuracy Assessment

Accuracy assessment is a method mainly used to enumerate the reliability of a classified image. In this standard procedure, construct a confusion matrix or error matrix to describe the performance of classified algorithms. The confusion matrix compares predicted values of a specific classifier with actual values of data then represents the correct performance of the classifier with the help of various evaluators.

In terms of four leading evaluators **(**Overall Accuracy (OA), Precision (P), Recall (R), and Misclassification rate (MR)**)**, the effectiveness of the proposed system was evaluated. Therefore, 3000 testing samples were selected for accuracy assessment. With the help of these testing samples and the confusion matrix of each model, we got accuracy assessment outcomes in Table 3. for two different classes.

Table 3. indicates a summary of classification accuracy obtained from supervised algorithms. The overall accuracy is 93% which shows better classification results for two different classes. Although each classifier formed good classification accuracy, KNN formed the highest accuracy compared to FR and maximum likelihood. M-LH produced minimum overall performance on a particular dataset. It means that there were high error rates and spectral similarities for the maximum likelihood classifier for the lake and surface of the glacier class.

Discussion

The monitoring of glacial lakes at a regional and global scale plays an essential role in lessening natural disasters, global warming, and various human threats (Yasmeen, 2013). In literature, different types of datasets and methodologies were used to monitor glacial lakes from mountains. The most conventional and commonly used dataset was optical data that produced the best results only in the absence of cloud cover. Although cloud cover or speckle noises, earlier research took more additional processing steps to eliminate noises and then reached the most satisfactory outcomes that may be time-consuming (Zhang, Chen, Tian, Liang, & Yang, 2019). The uniqueness of our study as a contrast to previous studies is that we used microwave GRD data in an alternative way that almost overcome the limitation of optical remote sensing images. This independent cloud data supports several filters in an effortless processing way to filter out speckle noises. The current study used those filters from which the mean, minimum, and maximum show the best results as they can have the ability to detect glacial lakes above as well as below the glacial surface. However, the other three parameters, such as standard deviation, maximum-

minimum ratio, and polarized ratio, have the lowest backscattering value for glacial lakes on the shallow of the glacier. The study (Liu, 2016) got the lowest backscatter value of standard deviation for open water bodies due to the specular reflectance over the smooth water surface.

Nevertheless, a recent study extraction of glacial lakes, more explicitly using GRD data, results from those obtained by (Wangchuk & Bolch, 2020). Still, it results in less period and easy implementation by microwave filters, unlike optical data. More innovatively, our proposed method used machine learning algorithms that improve the performance of remote sensing techniques with the help of classification. The implemented algorithms are K-nearest neighbor, Random Forest, and Maximum Likelihood, which gives a better understanding for extraction of glacial lakes at between 90–93% accuracy rate. Hence the proposed system is effective over the conventional method (Liu, 2016) due to integrating machine learning classifiers with backscatter analysis that is never done in preceding studies.

In mountainous areas, sometimes GLOFs occur due to increasing temperature and climate crisis(Mckinney, Byers, Rounce, Portocarrero, & Lamsal, 2015). So Hunza basin is one of the renowned mountainous regions along with numerous glaciers and glacial lakes. Therefore, various kind of precautions is necessary to mitigate the effects of GLOFs in such a region. Our study of glacial lakes helps reduce the negative impact on livelihood by giving future precautionary measurements of GLOFs events.

Conclusion

The main objective of this study is to analyze the importance of GRD images for the detection of glacial lake outlines. Since the Sentinel-1 GRD has a high spatial-temporal resolution, easily accessible, and independent of climate. Monitoring glacial lake outlines from high elevation mountains can reduce all the hazards related to glaciers and lakes by using SAR images. To meet the study objective, this research was done for the Passu Batura glacier of the Hunza basin, which is part of the Upper Indus Basin (UIB). The study manipulated two main approaches, polarized backscattering estimators and machine learning classifiers. In the first approach, the calculated values of mean, maximum, and minimum estimators behave well to distinguish lakes from the glacier's surface. So, these filters produced generally superior results for the study area. Besides, standard deviation and polarized ratio findings are unsatisfactory due to their comparable backscattering for lake and surface class. Similarly, the second approach was supervised machine learning algorithms. The three most commonly used ML algorithms were used to classify the GRD image into two classes (lake and glacial surface). The selected classifiers are FR, K-NN, and maximum likelihood. Each classifier produced predictions on the training dataset that were validated with actual values by using a confusion matrix. After the validation process, the results show that K-NN is comparatively more effective than maximum likelihood and RF classifiers. Finally, we conclude that ML algorithms and SAR backscattering analysis both techniques have overall higher outcomes in the study of glacier lakes. However, ML algorithms are considered core modules in the GIS and RS discipline as they enhanced the performance analysis of GRD data in a more reliable way.

Declarations

Acknowledgment

This research has been supported by the recurring research grant of Karakoram International University (KIU), Gilgit, 15100, Pakistan. We would like to acknowledge the Department of Computer Science faculty and supporting staff for their continued technical support and lab facility during the study period. We would like to extend our gratitude and thanks to anonymous reviewers for reviewing this manuscript.

Authors Contribution

All authors equally contributed from introduction to conclusion under the supervision of our principal investigator, and this task was completed with the collaboration of all authors. Mr. Syed Najam ul Hassan developed the main idea of research and supervised the study. Mrs. Hajra conducted this research and converted it into a manuscript. Dr. Garee Khan assisted in establishing graphs, maps, manuscript writing, help in modification and editing. Dr. Aftab Ahmad Khan made the provision of relevant literature and helped in the application of algorithms. Dr. Javed Akhter Qureshi reviews the MS before submission to the journal.

Competing Interest

Please refer to this article submitted to the **"Environmental Sciences and Pollution Research"** entitled **"Machine Learning Algorithms for Extraction of Glacial Lakes Using Ground Range Detected (GRD) Data: A Case Study from Hunza River Basin, Pakistan"** for publication. As the corresponding author, I undertake on behalf of myself and co-authors that:

Availability of data and materials: The datasets used or analyzedduring current study are available from the crossponding authors on reasonable request, and all the data generated or analyzed during this study are included in the published article.

Ethical Approval: Not Applicable

Consent to Participate: Not Applicable

Consent to Publish: Not Applicable

References

1. Ashraf, A., Naz, R., & Iqbal, M. B. (2017). Altitudinal dynamics of glacial lakes under changing climate in the Hindu Kush, Karakoram, and Himalaya ranges. *Geomorphology*, *283*, 72–79. https://doi.org/10.1016/j.geomorph.2017.01.033
2. Ashraf, A., Naz, R., & Roohi, R. (2012). *Glacial lake outburst flood hazards in Hindukush , Karakoram and Himalayan Ranges of Pakistan : implications and risk analysis*. *5705*.

https://doi.org/10.1080/19475705.2011.615344

3. Bazai, N. A., Cui, P., Carling, P. A., & Wang, H. (2020). Jo ur na l P re. *Earth-Science Reviews*, 103432. https://doi.org/10.1016/j.earscirev.2020.103432
4. Chen, F., Zhang, M., Tian, B., & Li, Z. (2017). *Extraction of Glacial Lake Outlines in Tibet Plateau Using Landsat 8 Imagery and Google Earth Engine*. *10*(9), 4002–4009.
5. Elsahabi, M. A., Negm, A. M., & Ali, K. A. (2016). *The Egyptian International Journal of Engineering Sciences and Technology Performances Evaluation of Water Body Extraction Techniques Using Landsat ETM + Data : Case- Study of Lake Nubia , Sudan*. *19*(2), 275–281.
6. *Glacial Lakes Have Grown Rapidly Worldwide- Satellite Images _ Earth*. (n.d.). Retrieved from https://earth.org/glacial-lakes-have-grown-rapidly-worldwide/
7. Gomathi, M., Geetha Priya, M., & Krishnaveni, D. (2018). Supervised classification for flood extent identification using sentinel-1 radar data. *Proceedings - 39th Asian Conference on Remote Sensing: Remote Sensing Enabling Prosperity, ACRS 2018*, *5*(October 2018), 3277–3284.
8. Hewitt, K. (2005). The Karakoram Anomaly? Glacier Expansion and the 'Elevation Effect,' Karakoram Himalaya. *Mountain Research and Development*, *25*(4), 332–340. https://doi.org/10.1659/0276-4741(2005)025[0332:tkagea]2.0.co;2
9. Khan, A. A., Jamil, A., Hussain, D., Taj, M., Jabeen, G., & Malik, M. K. (2020). Machine-Learning Algorithms for Mapping Debris-Covered Glaciers: The Hunza Basin Case Study. *IEEE Access*, *8*, 12725–12734. https://doi.org/10.1109/ACCESS.2020.2965768
10. Liu, C. (2016). Analysis of Sentinel-1 SAR data for mapping standing water in the Twente region. *University of Twente - ITC*. Retrieved from http://www.itc.nl/library/papers_2016/msc/wrem/cliu.pdf
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. https://doi.org/10.1080/01431161.2018.1433343
12. Mckinney, D. C., Byers, A. C., Rounce, D. R., Portocarrero, C., & Lamsal, D. (2015). *Assessing downstream flood impacts due to a potential GLOF from Imja Tsho in NepTshoaAssessing downstream flood impacts due to a potential GLOF from Imja in Nepal*. 1401–1412. https://doi.org/10.5194/hess-19-1401-2015
13. Mitkari, K. V., Arora, M. K., & Tiwari, R. K. (2017). Extraction of Glacial Lakes in Gangotri Glacier Using Object-Based Image Analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(12), 5275–5283. https://doi.org/10.1109/JSTARS.2017.2727506
14. Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine Learning Classification Techniques for Breast Cancer Diagnosis. *IOP Conference Series: Materials Science and Engineering*, *495*(1). https://doi.org/10.1088/1757-899X/495/1/012033
15. Onojeghuo, A. O., Blackburn, G. A., Wang, Q., Atkinson, P. M., Kindred, D., & Miao, Y. (2018). Mapping paddy rice fields by applying machine learning algorithms to multi-temporal sentinel-1A and landsat data. *International Journal of Remote Sensing*, *39*(4), 1042–1067. https://doi.org/10.1080/01431161.2017.1395969

16. *Second Summer School on Integrated Water Resources Management 27-31*. (2018).
17. Shugar, D. H., Burr, A., Haritashya, U. K., Kargel, J. S., Watson, C. S., Kennedy, M. C., … Strattman, K. (2020). Rapid worldwide growth of glacial lakes since 1990. *Nature Climate Change*, *10*(10), 939–945. https://doi.org/10.1038/s41558-020-0855-4
18. Verpoorter, C., Kutser, T., & Tranvik, L. (2012). Automated mapping of water bodies using landsat multispectral data. *Limnology and Oceanography: Methods*, *10*(DECEMBER), 1037–1050. https://doi.org/10.4319/lom.2012.10.1037
19. Wangchuk, S., & Bolch, T. (2020). Mapping of glacial lakes using Sentinel-1 and Sentinel-2 data and a random forest classifier: Strengths and challenges. *Science of Remote Sensing*, *2*(July), 100008. https://doi.org/10.1016/j.srs.2020.100008
20. Yao, X., Liu, S., Han, L., Sun, M., & Zhao, L. (2018). Definition and classification system of glacial lake for inventory and hazards study. *Journal of Geographical Sciences*, *28*(2), 193–205. https://doi.org/10.1007/s11442-018-1467-z
21. Yasmeen, Z. (2013). GLOF risk mapping in Hunza River Basin (Pakistan) using geospatial techniques. *Ieee*, 191–195.
22. Zhang, M., Chen, F., Tian, B., Liang, D., & Yang, A. (2019a). High-frequency glacial lake mapping using time series of Sentinel-1A/1B SAR imagery: An assessment for southeastern Tibetan Plateau. *Natural Hazards and Earth System Sciences Discussions*, (July), 1–18. https://doi.org/10.5194/nhess-2019-219
23. Zhang, M., Chen, F., Tian, B., Liang, D., & Yang, A. (2019b). High-frequency glacial lake mapping using time series of Sentinel-1A/1B SAR imagery: An assessment for southeastern Tibetan Plateau. *Natural Hazards and Earth System Sciences Discussions*, 1–18. https://doi.org/10.5194/nhess-2019-219

Figures

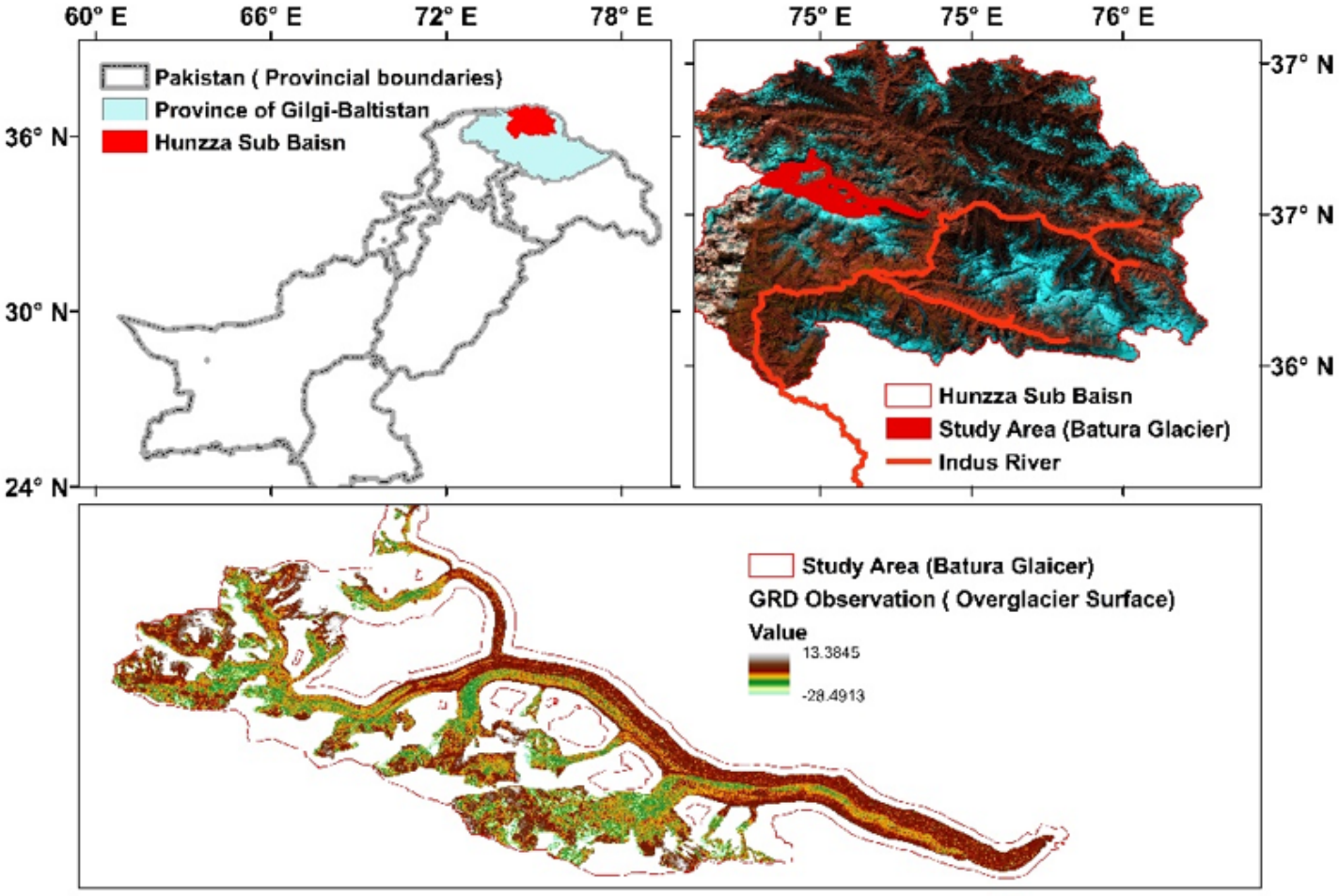

Figure 1

a) Pakistan map and locality of Hunza basin b) All possible glaciers and basic settlements of Hunza basin c) Batura selected as a study area

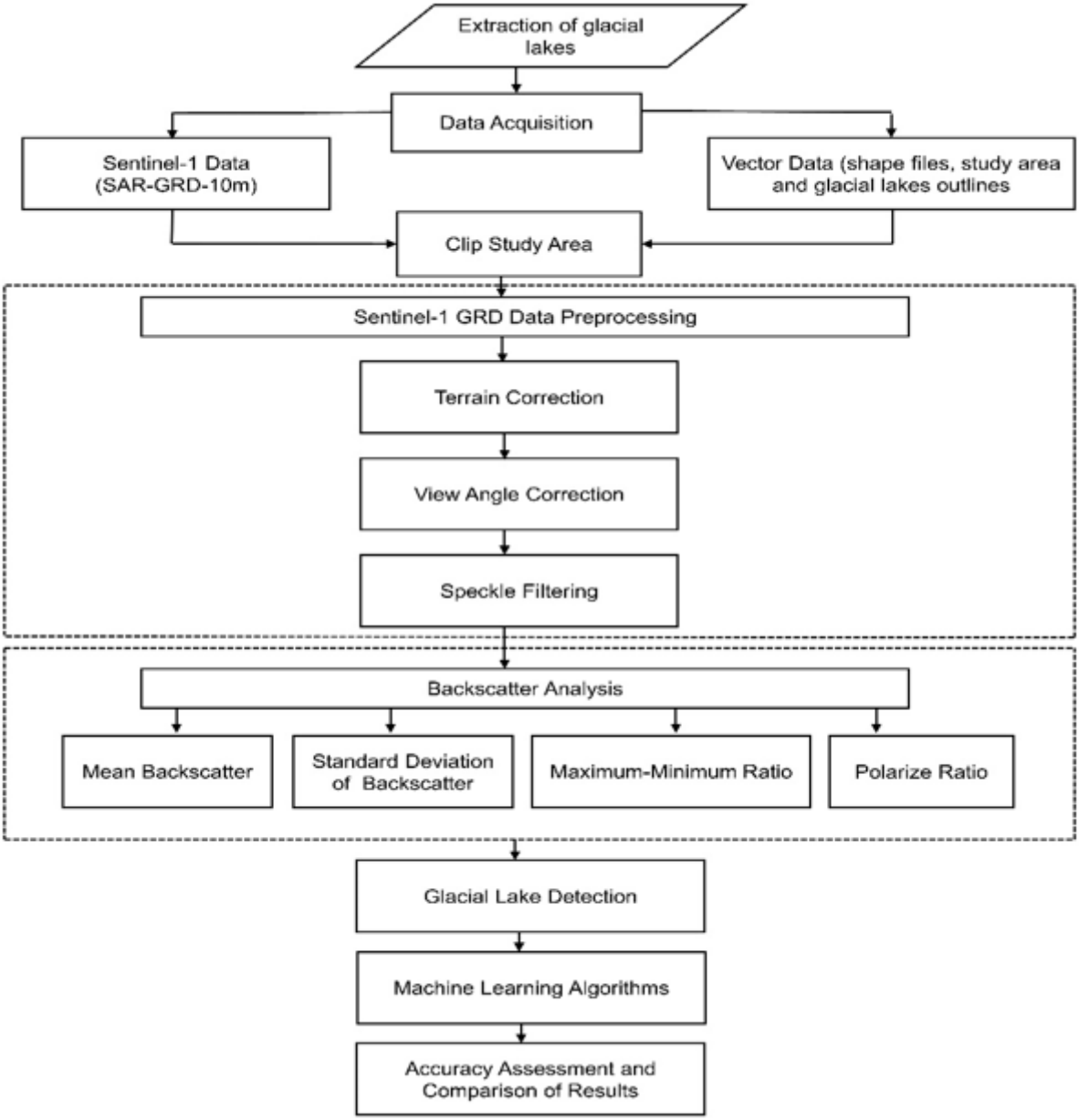

Figure 2

Methodology Flowchart Data Processing

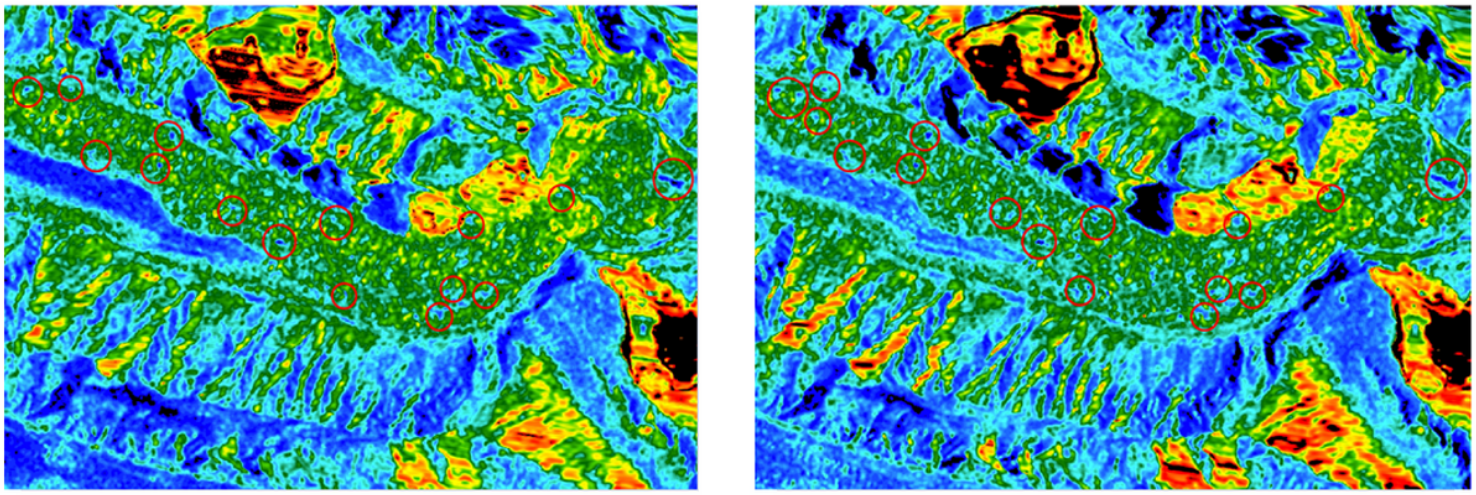

Figure 3

Backscatter polarized bands (a) Sigma0-VH (b) Sigma0-VV

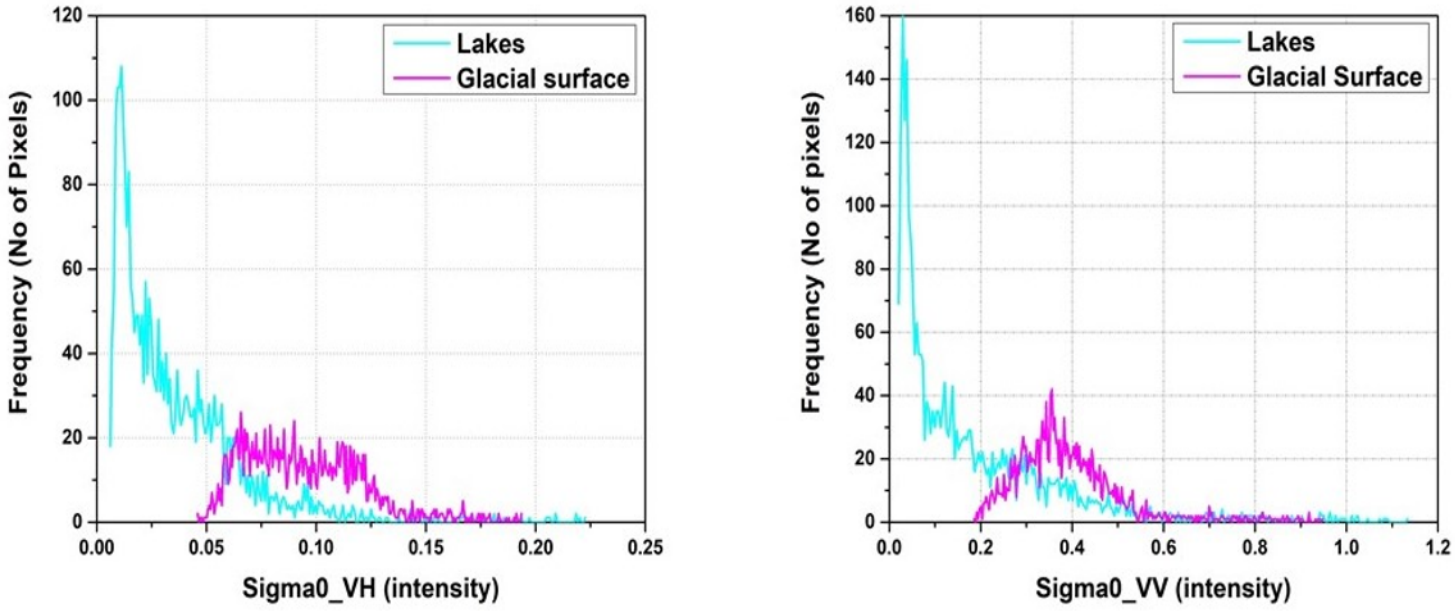

Figure 4

Performance of lakes and glacial surface in backscatter polarization

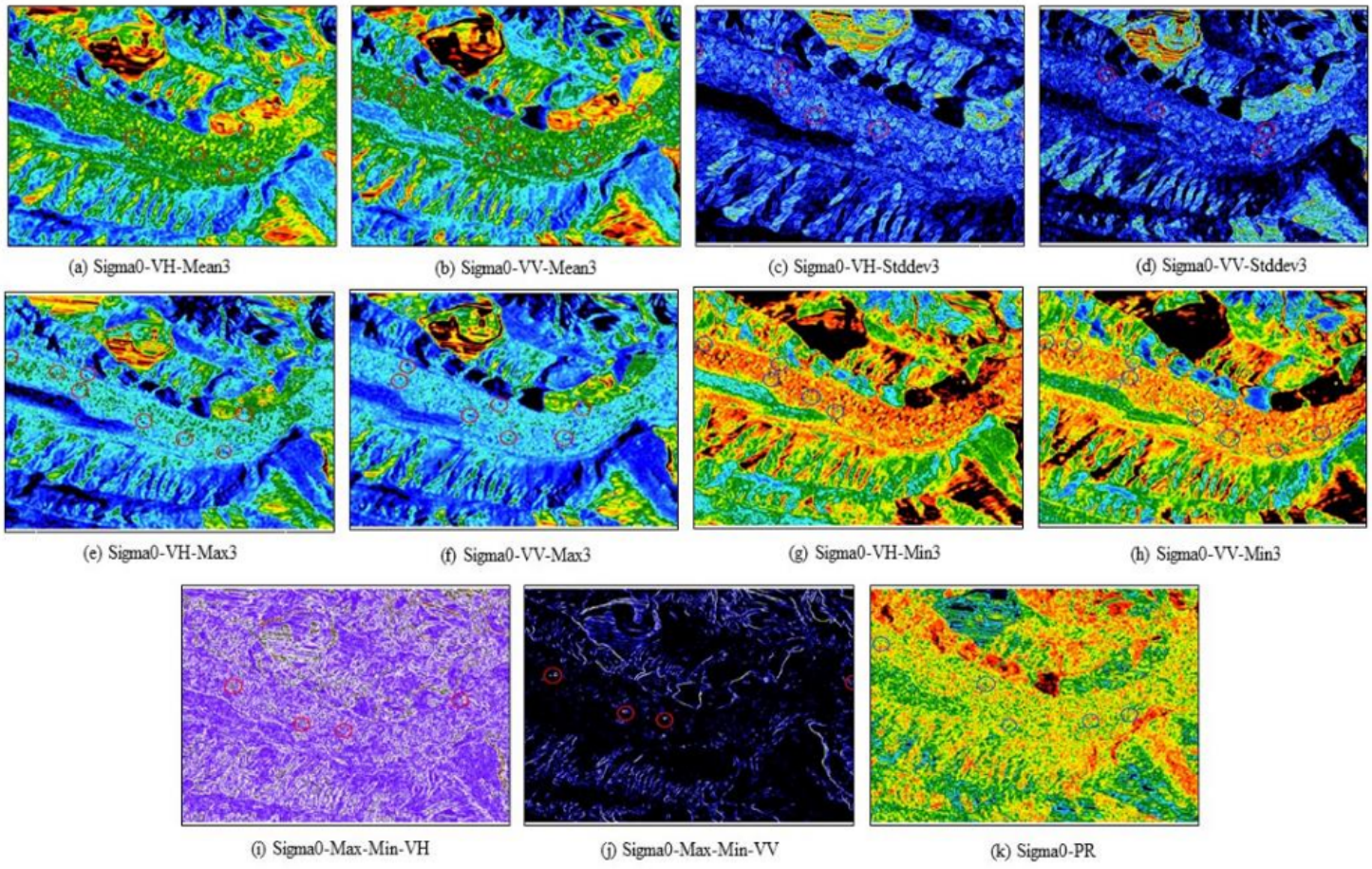

Figure 5

Map of backscatter estimators with RGB color composite Green-VH, Red, and Blue-VV polarization

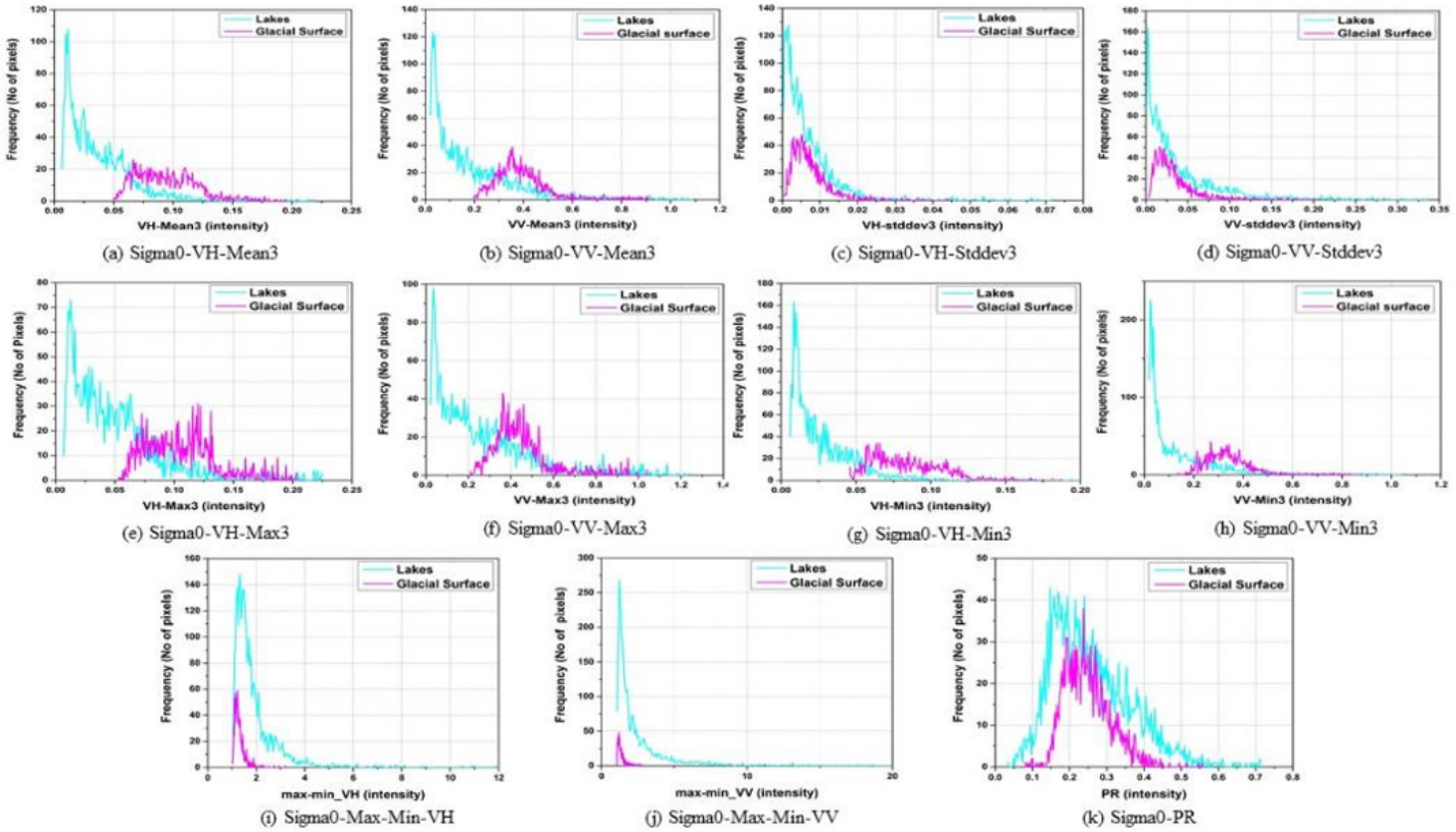

Figure 6

Graphs of backscattering estimators