2022

A Hybrid Machine Learning Technique For Feature Optimization In Object-Based Classification of Debris-Covered Glaciers

Shikha Sharda
I.K. Gujral Punjab Technical University, India

Mohit Srivastava
Chandigarh Engineering College, India

Hemendra Singh Gusain
Institute of Technology Management (ITM-DRDO), India

See next page for additional authors

Recommended Citation

Sharda, S., Srivastava, M. & Singh Gusain, H. (2022). A hybrid machine learning technique for feature optimization in object-based classification of debris-covered glaciers. *Ain Shams Engineering Journal*, vol. 13, pg. 101809. doi:10.1016/j.asej.2022.101809

Authors
Shikha Sharda, Mohit Srivastava, Hemendra Singh Gusain, Naveen Kumar Sharma, Kamaljit Singh Bhatia, Mohit Bajaj, Harsimrat Kaur, Hossam Zawbaa, and Salah Kamel

Ain Shams Engineering Journal 13 (2022) 101809

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

A hybrid machine learning technique for feature optimization in object-based classification of debris-covered glaciers

Shikha Sharda [a], Mohit Srivastava [b], Hemendra Singh Gusain [c], Naveen Kumar Sharma [d], Kamaljit Singh Bhatia [e], Mohit Bajaj [f], Harsimrat Kaur [g], Hossam M. Zawbaa [h,i,*], Salah Kamel [j]

[a] Department of Electronics and Communication Engineering, I.K. Gujral Punjab Technical University, Jalandhar, Punjab 144603, India
[b] Department of Electronics and Communication Engineering, Chandigarh Engineering College, Mohali, Punjab 140307, India
[c] Institute of Technology Management (ITM-DRDO), Mussoorie 248179, India
[d] Department of Electrical Engineering, I. K. Gujral Punjab Technical University, Jalandhar, Punjab 144603, India
[e] Department of Electronics and Communication Engineering, G.B. Pant Institute of Engineering and Technology, Pauri Garhwal 246194, India
[f] Department of Electrical and Electronics Engineering, National Institute of Technology, Delhi 110040, India
[g] Department of Electronics and Communication Engineering, CT Institute of Engineering and Technology, Jalandhar, Punjab 144623, India
[h] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
[i] Technological University Dublin, Dublin, Ireland
[j] Electrical Engineering Department, Faculty of Engineering, Aswan University, 81542 Aswan, Egypt

A R T I C L E I N F O

Article history:
Received 24 January 2022
Revised 12 March 2022
Accepted 5 April 2022
Available online 23 April 2022

Keywords:
Decision tree
Feature selection
Machine learning
Object-based classification

A B S T R A C T

Object-based features like spectral, topographic, and textural are supportive to determine debris-covered glacier classes. The original feature space includes relevant and irrelevant features. The inclusion of all these features increases the complexity and renders the classifier's performance. Therefore, feature space optimization is requisite for the classification process. Previous studies have shown a rigorous exercise in manually selecting the best combination of features to define the target class and proven to be a time-consuming task. The present study proposed a hybrid feature selection technique to automate the selection of the best suitable features. This study aimed to reduce the classifier's complexity and enhance the performance of the classification model. Relief-F and Pearson Correlation filter-based feature selection methods ranked features according to the relevance and filtered out irrelevant or less important features based on the defined condition. Later, the hybrid model selected the common features to get an optimal feature set. The proposed hybrid model was tested on Landsat 8 images of debris-covered glaciers in Central Karakoram Range and validated with present glacier inventories. The results showed that the classification accuracy of the proposed hybrid feature selection model with a Decision Tree classifier is 99.82%, which is better than the classification results obtained using other mapping techniques. In addition, the hybrid feature selection technique has sped up the process of classification by reducing the number of features by 77% without compromising the classification accuracy.

1. Introduction

Glaciers are moving mass of ice/snow with many other classes like debris, rocks, and surface lakes. Each of these classes has unique spectral behaviour and needs a specific combination of features to extract or classify them. This high-dimensional data contains hundreds of features comprising relevant and irrelevant feature sets. The inclusion of raw feature vectors leads to complex classification models, which results in poor performance in identifying class labels [1].

Various object-based prominent methods for classifying debris-covered glaciers have been introduced in the past few years [2–4].

* Corresponding author at: Technological University Dublin, Park House, 191 N Circular Rd, Cabra East, Grangegorman, Dublin D07 EWV4, Ireland.

E-mail addresses: rs.shikhasharda@ptu.ac.in (S. Sharda), mohit.ece@cgc.edu.in (M. Srivastava), mohitbajaj@nitdelhi.ac.in (M. Bajaj), hossam.zawbaa@gmail.com (H.M. Zawbaa), skamel@aswu.edu.eg (S. Kamel).

Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.asej.2022.101809
2090-4479/

These methods have used a combination of optical and topographic data for mapping lake, glacier ice, and debris-covered ice. However, these methods are semi-automatic. The authors examined various parameters like band ratios, spectral indices, topographic and spectral information to determine the most suitable features for the classification. Further, the threshold values to define glacier classes have been acquired either from previously published work or through the trial-and-error approach, which is a laborious task. Though, these existing methods have produced good classification accuracy but still require human efforts in selecting the best features and appropriate threshold values [2–4]. Therefore, determining how to automate the mapping of the debris-covered glacier with the effective use of available features needs profound research.

In machine learning, the feature selection methods are broadly categorized as filter, wrapper, and embedded. The filter methods compute the relevance of features based on statistical measures [5]. These methods are computationally less expensive and fast but do not always guarantee optimal results [5,6]. Relief-F, information gain, F-score, and correlation are the few widely used filter-based feature selection methods. The wrapper methods create several feature subsets and evaluate them using the predictive model. Each subset is assigned a score according to their performance. Based on these scores, an optimal feature subset is selected. Wrapper methods include simple greedy search algorithms and advanced evolutionary algorithms for feature selection [7]. These methods provide better results but are computationally more expensive, slower, and have a higher risk of over-fitting. The hybrid models combine filters and wrappers with taking advantage of both methods to overcome the drawbacks. The embedded methods select the optimal features based on the learning model like the random forest, classification and regression tree, and Lasso regression during the training phase. These methods are at low risk of over-fitting and computational less expensive than wrapper methods [5].

Many researchers have devised feature selection techniques to handle feature redundancy in biological [5,6,8], agricultural area [9–10], and urban area [11,12]; however, these techniques are yet to be established for glacier classification. The hybrid feature selection mechanism based on the filter and the wrapper methods has successfully improved the classification accuracy of biological problems [5,6,8]. Many multi-objective feature selection approaches have used particle swarm optimization [13,14], binary differential evolution algorithm [15], and genetic algorithm [16] to solve feature selection problems of standard datasets comprised of biological data, urban land cover data, and many more. Most wrapper methods and evolutionary algorithms are computationally infeasible for high-dimensional data [7]. The filter-based techniques have effectively optimized the feature space and improved the mapping accuracy of agricultural [10] and urban land cover areas [12] in an object-oriented environment. Bommert et al. [7] suggested that filter methods are computationally cheaper, faster, and can integrate with any machine learning model.

For glacier mapping, supervised machine learning classifiers have been introduced in the pixel-based environment to automate the extraction of debris-covered areas [17–19]. Thanki et al. [20] used a k-NN classifier for object-based mapping of glacier surfaces; however, this study mainly focused on mapping snow, debris-covered ice, lakes, vegetation, and rocks. Therefore, the existing object-based studies still lack the automatic mapping of debris-covered glaciers [21]. To date, object-based methods have not witnessed the use of feature selection techniques in extracting different land-cover classes of the debris-covered glacier. The feature space optimization before classification is highly significant for the effective use of high dimensional feature space comprising spectral, topographic, and textural information [1]. Therefore, considering the importance of feature space optimization and a need for an efficient automatic object-based mapping technique for debris-covered glaciers, the present study proposed the following objectives: (1) To introduce a hybrid feature selection technique that will reduce the classifier complexity and improve the prediction accuracy by automatically selecting the optimal feature set and eliminating the irrelevant/redundant features; and (2) To include a supervised machine learning-based classifier for automatic threshold parameters selection that will reduce the iterative attempts by the trial-and-error method in selecting appropriate threshold value for assigning objects to different classes.

Karakoram Range referred as a home to giant glaciers located at the western end of the Himalayas, has increasingly attracted researchers' attention because of its glaciological distinctiveness. Glacier mapping in Karakoram Range is challenging due to the debris-covered area. The previous studies in the Karakoram range mainly relied on the manual mapping of the debris-covered areas [22,23]. Therefore, the present research used the debris-covered glaciers over the Karakoram Range as a case study area to implement the proposed automatic object-based classification approach.

2. Related literature

The feature space optimization algorithms have gained popularity as they claimed to improve the classifier's performance and reduce the processing time by selecting a significant set of features [5,8]. This section reviewed different feature selection techniques used in object-based classification studies.

The Relief-F filter-based technique is considered one of the widely acceptable algorithms for multi-class problems [9]. Jia et al. [1] proposed an improved relief algorithm to optimize feature space for object-based classification of agricultural areas. This algorithm has reduced the original feature space by 67% and reported a significant increase of 6% in overall classification accuracy using optimized feature sets in comparison to classification using all features.

Ma et al. [10] evaluated different feature selection methods like filter, wrapper, and embedded to optimize the feature space for object-based classification of agricultural areas. Overall, the correlation-based feature selection method turned out to be the best feature subset evaluation method. Moreover, the feature subsets derived from wrapper methods have shown a negative impact on the performance of object-based classification. Whereas, results of the feature importance evaluation methods have positively affected the classifier's performance. This study demonstrated that the inclusion of feature importance evaluation methods prior to the classification step improves the classifier's performance.

Shi et al. [11] used a combination of Genetic Algorithm and Tabu Search (GATS) to select the optimal feature subset for classifying urban and suburban areas. The GATS method yielded the highest overall accuracy as compared to traditional GA, Relief-F, and multi-start Tabu search. However, the computational speed of GATS was the slowest among all.

Stromann et al. [12] analyzed the potential of three different dimensionality reduction methods like linear discriminant analysis (LDA), mutual information-Based (MI), and Fisher-criterion-based (F-score) in extracting optimal features for object-based classification of urban land cover classes. However, optimized feature sets obtained from filter-based feature selection techniques have improved the classification accuracy compared to the accuracy computed using an untreated feature set.

The above-mentioned studies acknowledge the benefit of prior feature space optimization in object-based classification and claim to improve classification accuracy. Also, the literature advised that the filter methods are valuable for object-based classification. Therefore, in the present study, filter-based feature selection methods: Relief-F and Pearson Correlation, were implemented and tested individually. However, these models failed in selecting appropriate feature sets to define land cover classes. Later, these models were combined to form a hybrid model and found satisfactory for the object-based classification of debris-covered glaciers.

3. Study area and dataset

Karakoram Range is a mountain range spanning over India and China borders, which covers a 500 km glacierized area in length. These mountains are the natural reservoir of frozen fresh water in the form of glaciers, ice sheets, and snow. The study area is a subregion on the Southern slope of the Central Karakoram with a geographic location between $76^{0}10'$E-$77^{0}15'$E and $35^{0}10'$N-$35^{0}50'$N (Fig. 1a). This area is heavily glaciated and contains the world's longest mountain glaciers outside the polar region like Siachen, Baltoro, and other small nearby glaciers like Kaberi, Chogolisa, and Sherpikang. Siachen glacier is stretching from north-northwest to south-southeast for 74 km. On the other hand, the Baltoro glacier extends over more than 60 km.

The Landsat 8 Operational Land Image and Thermal Infrared Sensor (OLI/TIRS) images of path/row: 148/35, acquired on 21 October 2020, having 30 m spatial resolution of the study area, was used to analyze the proposed technique as shown in Fig. 1(b-d). The whole study area divides into three regions: Region 1 corresponds to Siachen Glacier, Region 2 includes Kaberi, Chogolisa, and Sherpikang glaciers, and Region 3 covers Baltoro glacier. The satellite image was selected from the end of the glacier ablation period with 1.13% cloud cover to minimize the impact of seasonal snow cover for delineating glacier areas. These images were freely available from the EarthExplorer website developed by the United States Geological Survey (USGS).

3.1. Extracted features

Landsat 8 satellite has a total of eleven bands, including nine spectral bands (Coastal/Aerosol, Blue, Green, Red, Near-Infrared (NIR), Shortwave Infrared (SWIR1 and SWIR2), Panchromatic, Cirrus) and two thermal infrared bands (TIRS1 and TIRS2) and labelled in turn from 1 to 11. The Landsat 8 image was atmospherically cor-

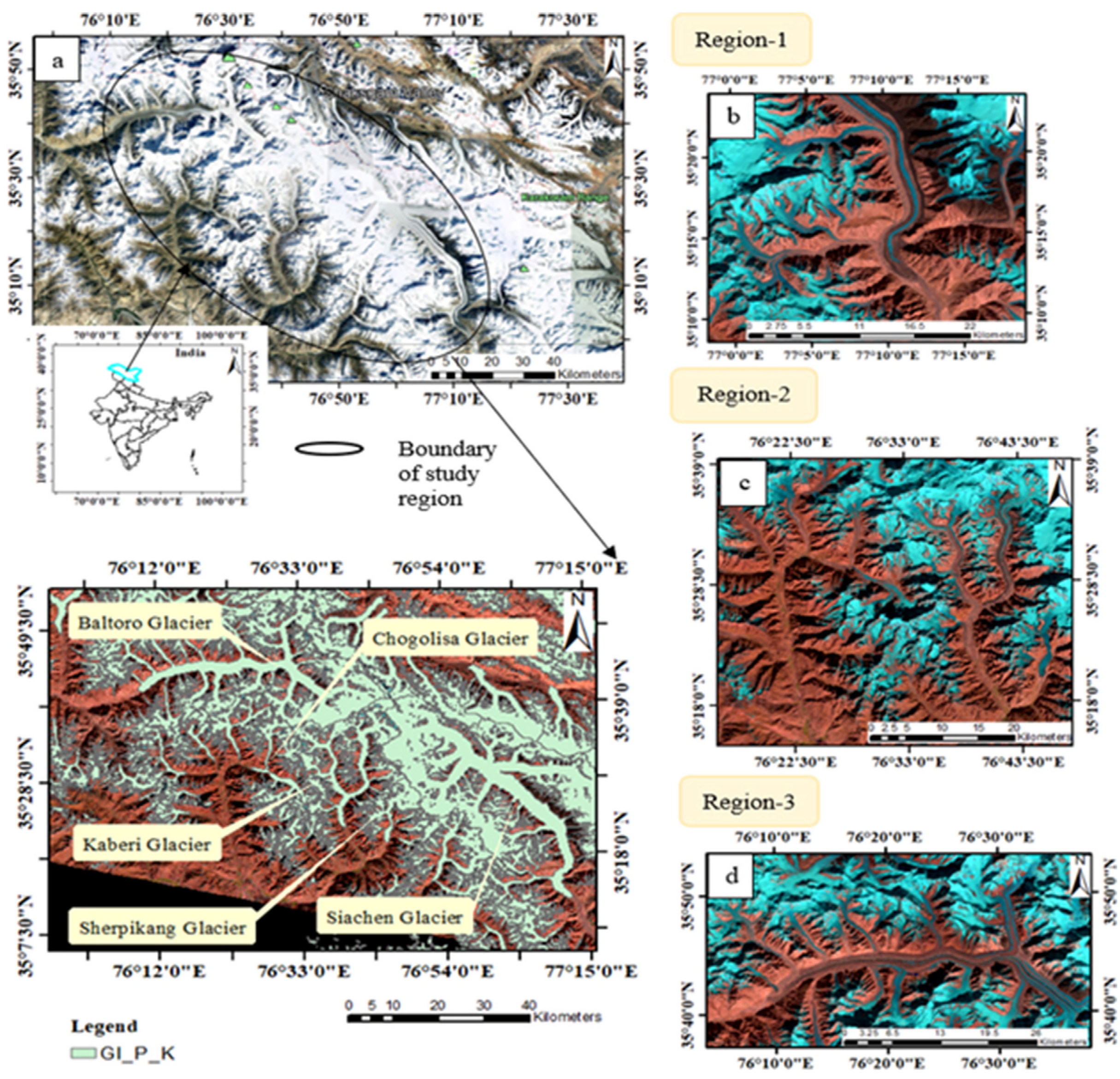

Fig. 1. (a) Location of the study area: Central Karakoram Range and (b-d) Landsat 8 image path/row:148/35 highlighting study areas. Image courtesy: Google Earth and USGS EarthExplorer.

rected using the Dark Object Subtraction method [24]. The spectral features include Mean and Standard Deviation for all eleven bands. The spectral index feature named Normalized Difference Snow Index (NDSI) [2] was extracted using Eq. (1) within eCognition Developer software.

$$NDSI = \frac{TM_{Green} - TM_{SWIR1}}{TM_{Green} + TM_{SWIR1}} \quad (1)$$

where TM_{Green} and TM_{SWIR1} are surface reflectance in Green and SWIR1 bands, respectively.

Textural features like Homogeneity, Contrast, Standard Deviation, Mean, Second Moment, and Correlation described by Haralick et al. [25] were extracted using eCognition Developer Software. These features were calculated using the Gray-Level Difference Vector (GLDV) and Gray-Level Co-Occurrence Matrix (GLCM).

Temperature and Topographic features are important parameters used in mapping debris-covered glaciers [17,19,26]. Therefore, the Land Surface Temperature (LST) was calculated from band10 of Landsat 8 as defined by Eq. (2) using ArcGIS Software (Fig. 2a).

$$LST = TB_{TOA}/1 + w * (TB_{TOA}/p) * \ln(LSE) \quad (2)$$

where TB_{TOA} is the top of atmosphere brightness temperature (°C) derived from TOA spectral radiance, w is the wavelength of emitted radiance (μm), p is a parameter derived from Planck's law, and LSE is the land surface emissivity computed from the proportion of vegetation.

Shuttle Radar Topography Mission 1-Arc Second Global Digital Elevation Model (SRTM-DEM), released by NASA and NGGA in cooperation with the German and Italian space agencies, had a 30 m resolution and was retrieved from EarthExplorer website for extracting slope information. Prior to analysis, the DEM data were projected to the same coordinate system as that of Landsat 8 image, i.e., World Geodetic System 1984 (WGS 84) with Universal Transverse Mercator (UTM) Zone 43 North. Then, the topographic feature slope was extracted from the SRTM DEM image using eCognition Developer Software (Fig. 2b).

3.2. Validation features

The ground truth data was not feasible due to the inaccessible terrain of glaciers. Therefore, two datasets: the Glacier inventory of the Pamir and Karakoram [27] and the Global Land Ice Measurements Space Initiative (GLIMS) glacier database [28], were applied to validate the glacier boundary detected by the proposed hybrid feature selection technique for object-based glacier mapping. Table 1 gives the detail of the dataset used in the study.

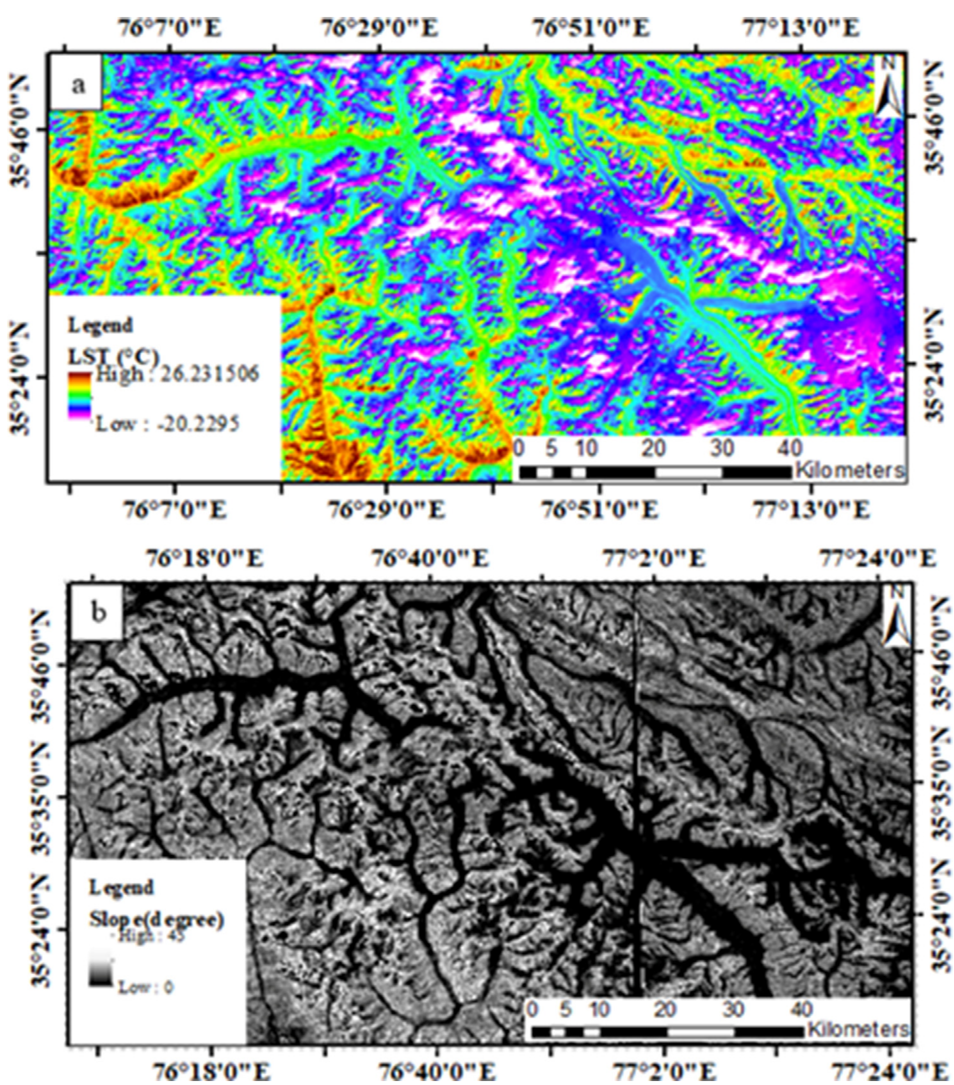

Fig. 2. (a) Land Surface Temperature (b) Slope maps for different land cover types of Central Karakoram glaciers.

Table 1
Detail of the datasets used in the study.

Datasets	Date	Resolution (m)
Landsat 8 (OLI & TIRS)	21 October 2020	30
SRTM 1-Arc Second GDEM	11 February 2000	30
Glacier inventory of the Pamir and Karakoram (GI_P_K)	2018	–
Global Land Ice Measurements Space Initiative (GLIMS) glacier database	2014	–

4. Analysis of features

Literature suggests spectral, thermal, and topographic data supports the mapping of debris-covered glaciers [19,29]. However, other features (textural and standard deviation) are used as auxiliary parameters and included in the initial feature space to prove their lower relevance by the proposed algorithm. Ice/snow cover has high reflectance in the visible bands and low reflectance in SWIR bands. Therefore, the spectral difference information from VIS-SWIR bands supports the mapping of snow/ice surfaces. The study region includes debris-covered glaciers, where debris has similar spectral characteristics to surrounding bare land in the Visible to Infrared band.

The LST map (Fig. 2a) demonstrates that supraglacial debris areas of Siachen, Baltoro, and Kaberi glaciers represented in green color have relatively lower surface temperatures than surrounding rocks. It is due to the freezing effect of beneath ice on supraglacial debris cover. Therefore, this temperature difference information is significant to separate them [29]. The Ice/snow cover mapped by the VIS-SWIR band ratio includes ice covered with debris. So, the spectral information in the NIR band can help separate ice mixed debris (IMD) area from clean ice/snow [17]. The selected study area is shadow-covered glaciers. The rocks in the shaded area have a lower temperature than non-shaded areas. Therefore, with the support of topographic slope feature debris area can be distinguished from the surrounding rocks in the shaded area. The black color in the slope map (Fig. 2b) represents the debris areas having a lower slope than the surrounding area highlighted in grey.

5. Proposed hybrid feature selection model: RF-Corr

Fig. 3 shows the procedure of the proposed hybrid feature selection model. The filter method used statistical measures to select feature subsets. Algorithm 1 describes the proposed hybrid feature selection technique, which includes three steps: initial screening, selection of common features, and refining. For Relief-F, initialized weight vector W to zero. Random 'm' training instances (R_i) out of total instances 'n' were selected. Distance information was computed by 'Manhattan' function to find the k nearest neighbours. The value of k = 3 was chosen that corresponds to maximum accuracy. The average feature weight W[A] of 3 nearest neighbours' instances (3 nearest Hits 'H' and Misses 'M'), was calculated as given in Eq. (3) [1]. Feature with the highest positive weight ranked at the top, and the rest arranged in descending order. Eq. (4) defines the difference function that measures the difference value of attribute A at two instances R_i and H_j or M_j. The minimum and maximum values of attribute A were selected over the entire selected instances.

Pearson correlation used correlation statistics to measure the degree of a linear relationship between the features and the targets. Features having low linear relations with other features and higher with class labels tend to be selected. The correlation was determined using Eq. (5). For optimal feature subset selection [30], the highest ranked features were initially included and computed the accuracy. The next ranked features were added singly and evaluated the performance. Few features when considered separately, may not be highly relevant but may perform well when combined with other features. Considering this, both models have selected the subset of features corresponding to maximum accuracy and less redundant. Features that have degraded the classifier's performance, were discarded and considered to be irrelevant. Next, the hybrid model inherits the properties of both models by selecting common features to maximize model efficiency further. The final relevant set of features as per the desired target was derived using Twoing split criteria as given in Eq. (6), which is offered by the Decision Tree (DT) classifier [31]. Rest, the default values of the parameters were used in building the tree. Twoing criteria check the probability of instances of a particular class appearing on the child node's left or right side. DT algorithm recursively partitions the training instances into subsets until all the subsets are assigned to a single class. The root node was selected based on the highest Twoing value of the feature.

Algorithm 1: Hybrid feature selection technique

Input: A vector of feature space X (n, a) values
 and class labels comprising:
a ← number of features/attributes
n ← number of instances
Output: A relevant set of features for classification
Initial Screening of Features
Relief-F algorithm:
A ← attribute/feature
m ← random training instances where m < n
P () ← probability
C ← corresponding class
for A = 1 to a **do**
Calculate weight vector

$$W[A] = W[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)/((m \times k)) + \sum_{C \neq class(R_i)} \Big[\frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C))\Big]/((m \times k)) \quad (3)$$

Where difference function

$$diff(A, R_i, H_j) = \frac{|value(A, R_i) - value(A, H_j)|}{\max(A) - min(A)} \quad (4)$$

end
return weight vector holding weight scores of all the features
 used to check their relevance
S_Features = X (:, rank(1:a));
Select top ranked features set that correspond to the
 maximum accuracy of the model
Pearson Correlation algorithm:
x ← instances
y ← class labels
Initialize correlation vector $Corr = 0$
for A = 1 to a **do**
Calculate correlation

$$Corr_{xy} = \frac{n\sum x_A y_A - \sum x_A \sum y_A}{\sqrt{n\sum x_A^2 - (\sum x_A)^2}\sqrt{n\sum y_A^2 - (\sum y_A)^2}} \quad (5)$$

(continued on next page)

Algorithm (*cotinues*)

end
Rank the features based on the correlation coefficient
S_Features = X (:, rank(1:a));
Select top ranked features set that correspond to the maximum accuracy of the model
Hybrid Model: RF-Corr
Input: $F_S = \{f_1, f_2, \ldots\ldots, f_s\} \leftarrow$ set of features from Relief-F
$F_N = \{f_1, f_2, \ldots\ldots, f_n\} \leftarrow$ set of features from Pearson Correlation
Output: Selection of Common Features, $F_{S \cap N} = F_S \cap F_N$
Final feature subset selection: Twoing Split Criteria (In-built function in DT-classifier)
Input: T ← Training instances
F ← Features set
Split criteria ← Attribute selection method
Output: dtree (T, F)
If stopping_cond is true **then**
Create a node N
If instances are of same class, C **then**
return N as a leaf node
else
create a root node
root.test_cond = find_best_split
x ← predictor variable
s ← chosen split at node t
i ← class
L(i), R(i) ← fraction of members of class i in the left child and right child respectively
$\delta(x,s,t)_{twoing} = P(L)P(R)/4(\sum_i |L(i) - R(i)|)^2$(6)
Assume, V = {v|v possible outcome of root.test_cond}
for each v ϵ V **do**
child = dtree (T, F)
Add child as descendent of root
end for
end if
return root

6. Classification results

The proposed feature selection algorithm was designed and implemented on MATLAB on a computer with the Windows 10 operating system and an intel core i5 processor.

6.1. Classification samples

The foremost step of object-based image classification was segmentation. The multiresolution image segmentation technique was employed to create homogeneous objects using eCognition developer software [32]. Total 74 object-based features were extracted along with these homogeneous segments, as summarized in Table 2. The initial feature space has been developed based on feature analysis from the available literature.

These objects/instances were assigned to five different land cover classes named Ice/snow, Supraglacial debris, Ice mixed debris (IMD), Terrain shadow, and Bare rock and sand based on visual inspection of Landsat 8 image, with the aid of high-resolution images of Google Earth to train the classifier (Fig. 4).

Later, all these features information of image objects were exported from eCognition developer software for further processing. The 5-fold cross-validation randomly split the dataset into a training/testing group (Table 2). Subsequently, 80% of the randomly selected instances form training samples, and the remaining 20% form testing samples to evaluate the classification results.

6.2. Experimental results

For Landsat 8 (OLI/TIRS), the total number of features was 74. In the initial screening of features, the Relief-F and Pearson Correlation models have selected the top-ranked features and filtered out the irrelevant ones having lower weight and correlation with respect to class labels. Table 3 represents the results of optimized feature space obtained by the individual and hybrid models. The number of features is reduced from 74 to 17 for Relief-F and 22 for Pearson Correlation. Later, the common features were selected using a hybrid RF-Corr model that has reduced the number of features by 77% (74 → 17).

Relief-F and Pearson Correlation models have ranked spectral index, mean and standard deviation features of individual spectral bands, LST, and slope features higher than textural features. Textural features are mainly required to examine surface characteristics (smoothness/coarseness, roughness), which is beyond the scope of the current study; hence, given least importance with respect to defined classes. Later, a subset of features that corresponds to maximum accuracy was selected. The optimized feature sets obtained from individual models named Relief-F, and Pearson Correlation was fed to DT classifier for final tunning by Twoing criteria [33]. Out of 17 and 22 features, DT has selected NDSI for ice/snow, Brightness for Terrain shadow, Slope followed by Mean TIRS to separate supraglacial debris-covered area, and Mean NIR for MID over the other spectral features. However, the TIRS feature resulted in a few misclassifications while mapping debris area, as represented in Fig. 5.

Next, the hybrid model selected the common features from the optimized feature sets of Relief-F and Pearson Correlation. The hybrid model selected the same features as that by the Relief-F model. However, DT has performed well with this optimized feature set by selecting all the desirable features (NDSI, Brightness, Slope, LST, and Mean (5)) required for mapping debris-covered glaciers. The classified map (Fig. 5) shows that the LST feature has effectively mapped the debris area boundary in contrast to the TIRS feature. Moreover, the existing studies have also considered the LST as the essential feature for the classification of supraglacial debris [17,19]. The final feature set for all the three regions obtained by the DT classifier using the hybrid feature selection model was in accordance with the existing debris-covered glacier mapping studies; hence, making the proposed technique more reliable. Table 4 summarizes the final feature set with automatically generated threshold ranges used for classifying different land cover classes: Ice/snow, Supraglacial debris, IMD, Terrain shadow, and Bare rock. A threshold value of 0.4 for NDSI, generally used in literature, as effective in mapping clean ice. Lu et al. [19] have used the 0.4 threshold value of NDSI to map snow/ice cover, resulting in incorrect classification of some areas. Later, they have manually reclassified the incorrectly classified areas (shadowed ice pixels) to snow/ice class. The acceptable threshold value of NDSI for mapping snow has been found in the range of 0.25 to 0.45 [19,34]. Moreover, the present study also confirmed the threshold value of NDSI between 0.25 and 0.28, as ideal for mapping ice/snow cover over Central Karakoram Range. The threshold values vary with variations in glacier topography.

Initially, the mapped ice/snow cover includes IMD. In the next step, Slope and Mean NIR features have separated IMD area from snow/ice cover. The LST value in the range of 14.01–15.78 °C, was found suitable for mapping supraglacial debris [17]. Fig. 6 represents the final glacier classified maps obtained using the proposed hybrid feature selection technique. The classified map states that the selected features and threshold values are appropriate for mapping debris-covered glaciers in the Central Karakoram range.

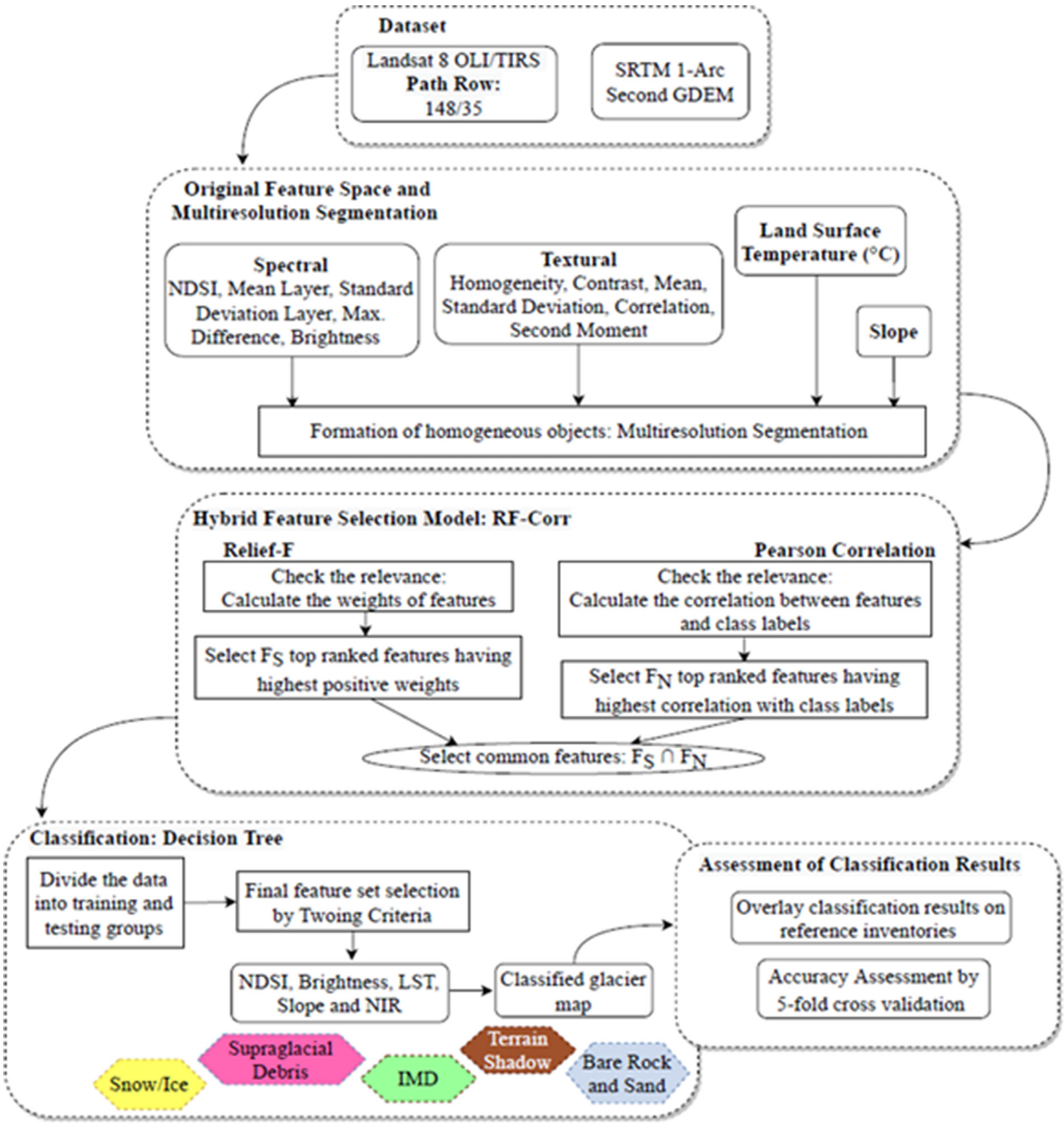

Fig. 3. Flowchart of the hybrid feature selection mechanism for automatic object-based glacier mapping.

Table 2
Description of the data.

Datasets	Attributes/Features	Instances/Objects	Class	Training/Testing group
Region 1	*Spectral* - Brightness, Maximum difference, NDSI, Mean layer (1/2/3/4/5/6/7/8/9/10/11),	2502	5	2002/500
Region 2	Standard Deviation (1/2/3/4/5/6/7/8/9/10/11)	2871	5	2297/574
Region 3	*Temperature* - LST	2325	5	1860/465
	Topographic - Slope			
	Texture - Texture by Haralick			
	GLCM & GLDV for Homogeneity, Contrast, Standard Deviation, Mean, Moment, Correlation			

7. Accuracy assessment

The effectiveness of the proposed hybrid feature selection (FS) technique was evaluated by calculating prediction accuracy using 5-fold cross-validation and the accuracy results were compared with the individual model's performance in Table 5. In comparison to Relief-F and Pearson Correlation models, the proposed hybrid model has slightly improved the classification accuracy by 0.04% in the case of the Siachen glacier and 0.17% in the case of other glaciers. Moreover, the DT classifier failed to choose a suitable combination of features from the optimized feature set generated by individual models [5]. Hence, does not recommend for future

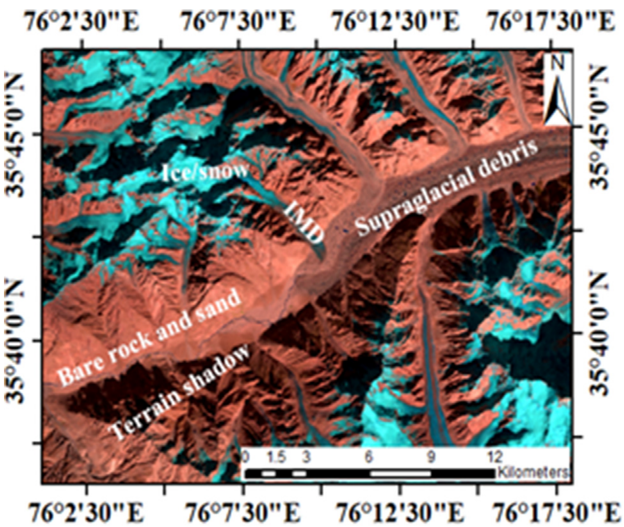

Fig. 4. Highlight five different land cover classes on the false-color composite (Bands: 6/5/4) of Landsat 8 image.

analysis. However, a combination of the hybrid model and DT classifier has performed satisfactorily for all three glacier regions.

The Confusion Matrix of the DT classifier using the proposed hybrid technique and overall Kappa statistics for 'Region 1', is represented to provide more detail about classification results (Table 6). For Region 1, the Kappa coefficient value was 0.993, which was within the acceptable range [17].

The areas of different land cover types, mapped by the proposed classification technique, were estimated within eCognition Developer software (Table 7). The whole study region was dominated by glacierized area (including ice/snow, supraglacial debris, and ice mixed debris) that covers 50.53% of the total area.

8. Discussion

8.1. Uncertainties of glacier data

There were many uncertainties during the process of mapping debris-covered glaciers by the proposed automatic technique. First, the present study used a Landsat 8 image having a spatial resolution of 30 m to map debris-covered glaciers, where it was hard to identify the small features. However, the higher-resolution Sentinel series satellite images can help in identifying the small features. Second, uncertainty was because of the training samples of the study area, selected based on the visual inspection of the

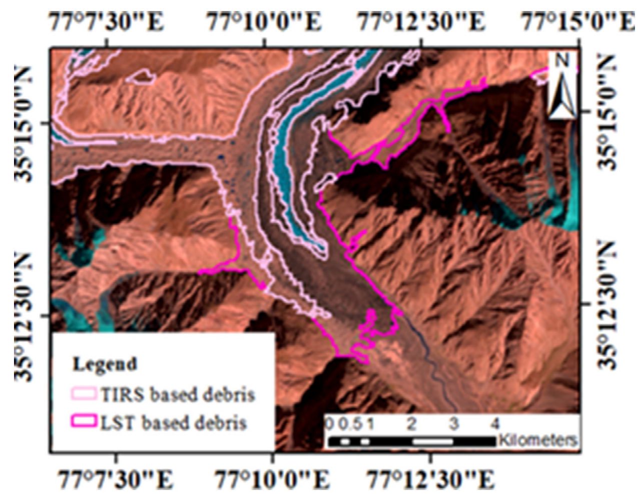

Fig. 5. Supraglacial debris area mapped by TIRS and LST features.

Table 4
Selected feature set with threshold values used to define land cover classes.

Land cover	Features with threshold condition range
Ice/snow	NDSI ≥ (0.25 – 0.28)
Supraglacial debris	Slope < (6.48^0 – 10.67^0) and LST < (14.01 °C– 15.78 °C)
Ice mixed debris (IMD)	Slope < (8.99^0 – 10.67^0) and NIR < (17979–19543)
Terrain shadow	Slope ≥ (15.03^0 – 15.41^0) and Brightness < (8999.19 – 10987.8)
Bare rock and sand	NDSI < (0.25 – 0.28) and Brightness ≥ (8999.19 – 10987.8)

Landsat 8 image with the aid of high-resolution images of Google Earth. However, the high-resolution images from Google Earth in the present study area were usually acquired in the winter, making it hard to identify debris-covered areas due to heavy snow cover. Therefore, a comprehensive field survey could help recognize the pixels as debris-covered areas or other land cover types of the glacier. Hence, an in-situ understanding of glacier surface area will be helpful for the precise mapping of glaciers.

Furthermore, DEM accuracy was a crucial factor in mapping debris-covered glaciers, which affected the topographic features of glaciers. The SRTM DEM data was acquired on 11 February 2000 and available online on 23 September 2014. The difference in the release date of the SRTM DEM image and the acquisition date of the Landsat 8 image might cause uncertainties in mapping debris areas. Although, this period recorded no visible elevation changes in the Southern slope of the Central Karakoram Range.

Table 3
Screening of features by different feature selection models and Twoing criteria for all the three study regions.

Feature selection algorithm	Optimized feature space	Final feature set selected by Twoing split criteria within Decision Tree
Relief-F (17)	Spectral – NDSI, Brightness, Maximum Difference, Mean Layer (1/2/3/4/5/6/7/8/10/11), Standard Deviation (6/7) Temperature – LST Topographic – Slope	NDSI, Brightness, Slope, Mean (11/5)
Pearson Correlation (22)	Spectral – NDSI, Brightness, Maximum Difference, Mean Layer (1/2/3/4/5/6/7/8/10/11), Standard Deviation (1/2/3/4/6/7/8) Temperature – LST Topographic – Slope	NDSI, Brightness, Slope, Mean (11/5)
Hybrid RF-Corr (17)	Spectral – NDSI, Brightness, Maximum Difference, Mean Layer (1/2/3/4/5/6/7/8/10/11), Standard Deviation (6/7) Temperature – LST Topographic – Slope	NDSI, Brightness, Slope, LST, Mean (5)

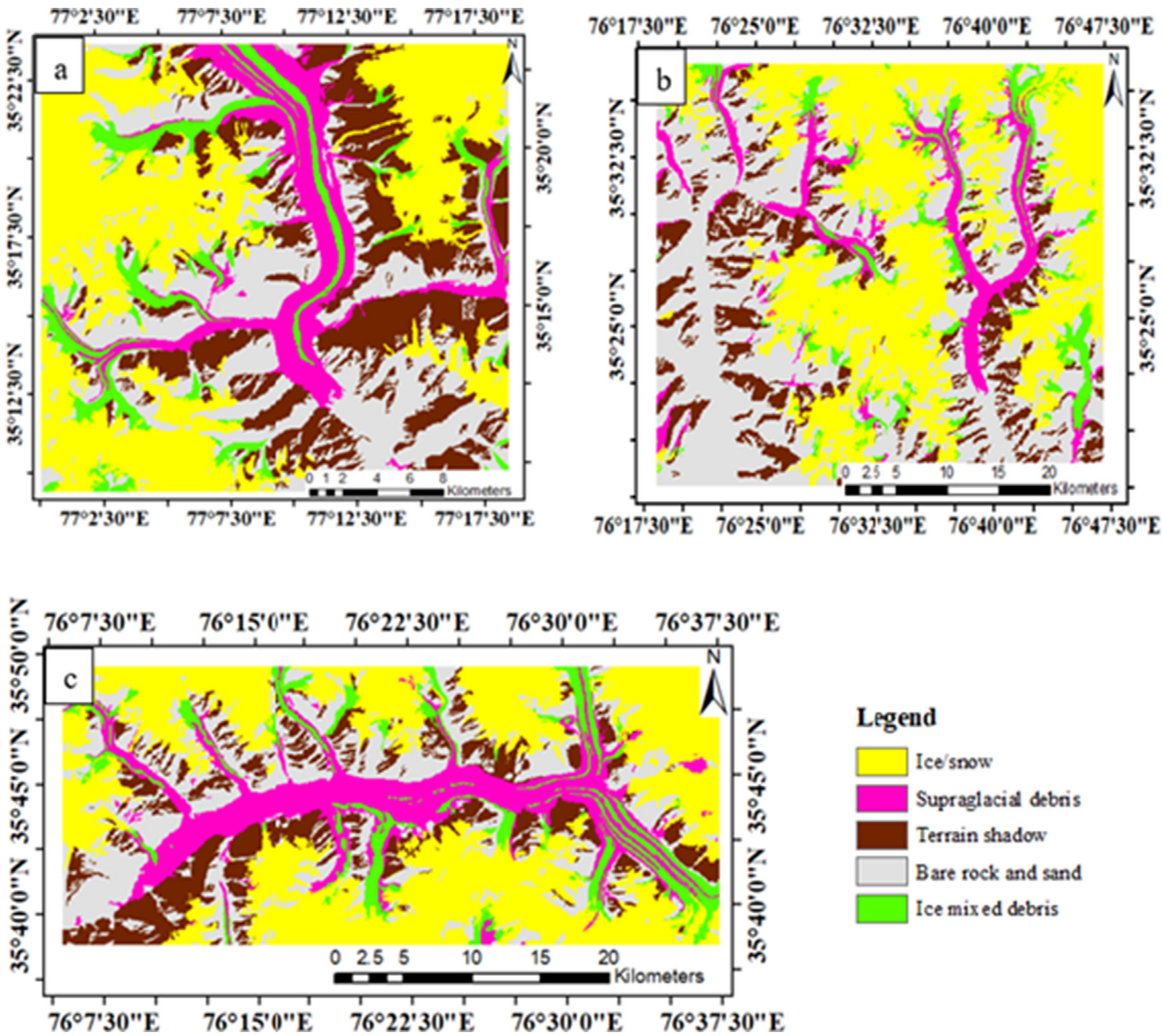

Fig. 6. Classification results of Landsat 8 images for (a) Region 1 (b) Region 2 (c) Region 3 using the proposed glacier mapping technique.

Table 5
K-Fold cross-validation based accuracy assessment (%) of different feature selection models.

Datasets	Relief-F	Pearson Correlation	Hybrid RF-Corr
Region 1	99.76	99.76	99.80
Region 2	99.65	99.65	99.82
Region 3	99.65	99.65	99.82

Table 6
Confusion matrix of proposed glacier classification technique for testing samples of region 1.

Land cover class	Ice/snow	Supraglacial debris	Ice mixed debris	Terrain shadow	Bare rock and sand	Total
Ice/snow	237	0	0	0	0	237
Supraglacial debris	0	14	0	0	0	14
Ice mixed debris	1	0	33	0	0	34
Terrain shadow	0	0	0	76	0	76
Bare rock and sand	0	0	0	0	139	139
Total	238	14	33	76	139	
Overall Kappa coefficient: 0.993						

However, it is essential to match the time frame with the dataset. Moreover, the thermal infrared band of Landsat 8 mapped the glaciers based on the assumption that the debris areas have a lower temperature than the surrounding rock surface [19], which is not applicable for highly thick-debris covered glaciers [35]. The ground truth data could help in the understanding of these glacier types.

8.2. Comparison with existing glacier mapping techniques

Previously, studies related to glacier mass change have been reported in the Central Karakoram range [22,23]. They have utilized the thresholding band ratio technique for mapping glacier ice region. These existing techniques are semi-automatic that needs a

Table 7
Area (Km^2) for different land cover classes mapped by the proposed technique.

Datasets	Ice/snow with IMD	Supraglacial debris	Terrain shadow	Bare rock and sand	Total
Region 1	295.43	70.94	187.03	183.37	735.30
Region 2	558.45	69.46	330.11	538.46	1496.48
Region 3	453.27	128.48	206.44	160.81	949

manual selection of thresholds. Moreover, the debris-covered area of Siachen and Baltoro glaciers was delineated manually using spectral bands of Landsat image [22,23]. The manual approach demands a high workload and their mapping accuracy counts on the expert's skill in identifying glacier outlines. The GLIMS database has used a semi-automatic object-based technique for extracting the glacier boundaries in the Central Karakoram range [28].

Haireti et al. [29] and Lv et al. [36] utilized semi-automatic pixel-based technique to delineate glaciers in the Eastern Karakoram range. These studies indicate the human dependency in selecting appropriate features and thresholding band ratios that make the classification a time-consuming process. Moreover, the current object-based glacier mapping studies still need manual assistance in mapping debris-covered glacier surfaces [2–4]. However, the proposed machine-learning-based classification technique has expedited the mapping of debris-covered glaciers due to its automatic selection of relevant features and estimation of appropriate threshold values; hence, making it reliable for mapping the large glacial regions.

Table 8 compares the performance of the proposed classification technique with the existing supervised object-based glacier mapping techniques to prove its usefulness. The k-Nearest Neighbor (k-NN) algorithm [20] was tested on the present study region for mapping debris-covered areas and found nearly 91% accurate. Next, the inclusion of the proposed hybrid FS technique prior to the k-NN classifier has sped the classification process by 2.8 sec with marginal improvement in performance accuracy. Another supervised classifier, Decision Tree having the inbuilt capability of estimating feature importance, was implemented and examined [37]. Though DT provides feature selection, its performance accuracy is less as compared to the proposed technique. The inclusion of the feature selection model before the classification process has accelerated the classification process by reducing the features by 77% (Table 8). The classifier takes less time to process the reduced feature space in comparison to the original feature space.

Other accuracy assessment statistical parameters associated with confusion matrices obtained using proposed and existing techniques for all three study regions were summarized in Table 9. The overall Kappa coefficient value of the proposed technique is greater than 0.8, which signifies the good quality of classification results for all land cover types. The overall Error rate reported in mapping land cover areas using the proposed technique was less as compared to the existing approach. In summary, the proposed classification technique has outperformed the existing classification techniques.

8.3. Comparison with glacier inventories

The classification results obtained by the DT classifier using the proposed hybrid feature selection technique were compared with two existing glacier inventories: the glacier inventory of the Pamir and Karakoram (GI_P_K) [27] and the Global Land Ice Measurements Space Initiative (GLIMS) [28] glacier database. GI_P_K is the homogenous inventory of Pamir and Karakoram mountains formed by analyzing 28 Landsat 4–5 Thematic Mapper (TM) and Landsat 7 Enhanced Thematic Mapper (ETM+) images acquired from 1998 to 2002. GI_P_K applied well-established methods of mapping glaciers and utilized the coherence maps obtained from Advanced Land Observing Satellite 1 (ALOS-1) Phased Array type L-band Synthetic Aperture Radar 1 (PALSAR-1) to support the glaciers mapping. GLIMS provides glacier boundaries covering debris-covered regions across the globe. Landsat ETM+ and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) scenes were utilized to prepare the GLIMS database and maintained at National Snow and Ice Data Center (NSIDC). The International Centre for Integrated Mountain Development (ICIMOD), a Regional Center of GLIMS, has analyzed the Landsat images of the Himalaya Mountains using a semi-automatic object-based technique. The GI_P_K and ICIMOD inventories have used Landsat ETM+ scenes from 2001 and 2006, respectively, for the selected study area.

Fig. 7 shows the results of the proposed classification technique overlaid on two glacier inventories. The composite map represents some differences in glacier outlines that are likely to be caused due

Table 8
Comparison of the proposed technique with existing supervised object-based approach [20] based on K-fold cross-validation accuracy assessment (%) and computational time (seconds).

Datasets	K-fold cross-validation accuracy assessment (%)				Computational time (seconds)			
	k-NN	Hybrid FS & k-NN	DT	Proposed technique (Hybrid FS & DT)	k-NN	Hybrid FS & k-NN	DT	Proposed technique (Hybrid FS & DT)
Region 1	91.40	91.40	99.72	99.80	3.05	0.19	3.58	0.38
Region 2	91.60	91.63	99.47	99.82	3.15	0.25	4.07	0.67
Region 3	90.66	90.96	99.69	99.82	3.08	0.21	3.52	0.32

Table 9
Overall accuracy assessment statistical parameters based on confusion matrices for proposed technique and existing supervised object-based approach [20].

Datasets	Error rate (%)		Kappa coefficient	
	k-NN	Proposed technique	k-NN	Proposed technique
Region 1	8	0.2	0.75	0.993
Region 2	7.3	0.8	0.771	0.972
Region 3	8.82	0.4	0.724	0.986

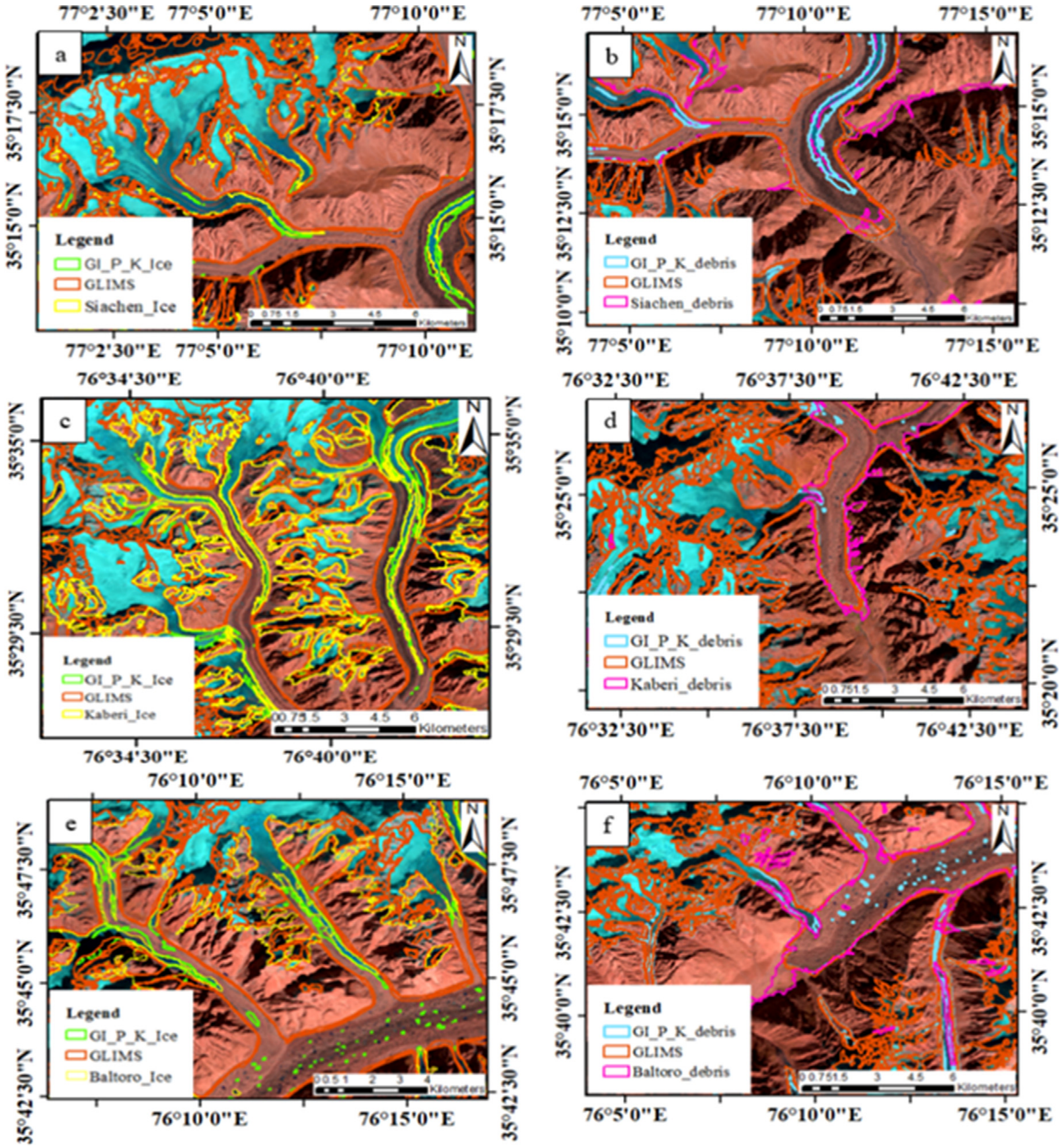

Fig. 7. (a-f) Comparison of classification results of proposed glacier mapping technique with two datasets: Glacier inventory of the Pamir and Karakoram (GI_P_K) and Global Land Ice Measurements Space Initiative (GLIMS) glacier database for different glaciers.

to differences in the acquisition time of images used in the study. The boundary of the glacier ice region obtained by the proposed technique nearly fitted with the glacier boundary of both the inventories. Small parts show minor differences in the outline of the ice boundary could be due to seasonal snow. The edges of the debris-covered region mapped by the proposed technique were slightly different compared to inventories. Overall, the glacier area outline shows a high degree of agreement with the previous datasets making it robust for mapping debris-covered glaciers.

9. Conclusion

In object-based debris-covered glacier classification, feature space optimization is a prerequisite step to be performed before the classification process. However, the inclusion of irrelevant features increases classifier complexity and degrades the mapping accuracy. The present research has introduced a hybrid feature selection technique comprising three steps: initial screening, selection of common features, and refining. The combination of Relief-F and Pearson Correlation filter-based techniques has optimized the feature space. DT classifier has further refined the optimized feature space by using Twoing split criteria. The proposed machine-learning-based automatic classification technique was tested in Central Karakoram Region and has shown high robustness in all the glaciers. The combination of feature selection technique and DT algorithm proves to be advantageous in eliminating the irrelevant/redundant features, reducing classifier complexities, reducing the computational time and enhancing the classification accuracy. This automatic approach of object-based mapping debris-covered glaciers will benefit glaciologists and researchers working on the scientific study of glaciers and provide technical assistance in developing future glacier inventory since it has greatly reduced human efforts. In the future, the proposed technique can successfully be utilized for mapping glaciers in some other areas using high-resolution Sentinel-2 satellite data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank I.K. Gujral Punjab Technical University, Jalandhar, Punjab, India, for providing the support to carry out this study. We are grateful to USGS and NASA for providing the Landsat 8 and SRTM DEM satellite data (freely available at https://earthexplorer.usgs.gov).

The author (Hossam M. Zawbaa) thanks the European Union's Horizon 2020 research and Enterprise Ireland for their support under the Marie Skłodowska-Curie grant agreement No. 847402.

References

[1] Jia J, Yang N, Zhang C, Yue A, Yang J, Zhu D. Object-oriented feature selection of high spatial resolution images using an improved relief algorithm. Math Comput Model 2013;58(3–4):619–26. doi: https://doi.org/10.1016/j.mcm.2011.10.045.
[2] Robson BA, Nuth C, Dahl SO, Hölbling D, Strozzi T, Nielsen PR. Automated classification of debris-covered glaciers combining optical, SAR and topographic data in an object-based environment. Remote Sens Environ 2015;170:372–87. doi: https://doi.org/10.1016/j.rse.2015.10.001.
[3] Rastner P, Bolch T, Notarnicola C, Paul F. A comparison of pixel- and object-based glacier classification with optical satellite images. IEEE J Sel Top Appl Earth Obs Remote Sens 2014;7(3):853–62.
[4] Muhammad S, Gul C, Javed A, Muneer J, Waqar MM. Comparison of glacier change detection using pixel based and object-based classification techniques. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Melbourne, 2013.
[5] Thejas GS, Garg R, Iyengar SS, Sunitha NR, Badrinath P, Chennupati S. Metric and accuracy ranked feature inclusion: Hybrids of filter and wrapper feature selection approaches. IEEE Access 2021;9:128687–701. doi: https://doi.org/10.1109/ACCESS.2021.3112169.
[6] Hsu HH, Hsieh CW, Lu MD. Hybrid feature selection by combining filters and wrappers. Expert Syst Appl 2011;38(7):8144–50. doi: https://doi.org/10.1016/j.eswa.2010.12.156.
[7] Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal 2020;143:106839.
[8] Mohamed AAA, Hassan SA, Hemeida AM, Alkhalaf S, Mahmoud MMM, Eldin AMB. Parasitism – Predation algorithm (PPA): A novel approach for feature selection. Ain Shams Eng J 2020;11(2):293–308. doi: https://doi.org/10.1016/j.asej.2019.10.004.
[9] Kiala Z, Mutanga O, Odindi J, Peerbhay K. Feature selection on Sentinel-2 multispectral imagery for mapping a landscape infested by parthenium weed. Remote Sens 2019;11(16):1892. doi: https://doi.org/10.3390/rs11161892.
[10] Ma L, Fu T, Blaschke T, Li M, Tiede D, Zhou Z, et al. Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. ISPRS Int J Geo-Inf 2017;6(2):51. doi: https://doi.org/10.3390/ijgi6020051.
[11] Shi L, Wan Y, Gao X, Wang M. Feature selection for object-based classification of high-resolution remote sensing images based on the combination of a genetic algorithm and tabu search. Comput Intell Neurosci 2018;2018:1–13.
[12] Stromann O, Nascetti A, Yousif O, Ban Y. Dimensionality reduction and feature selection for object-based land cover classification based on Sentinel-1 and Sentinel-2 time series using google earth engine. Remote Sens 2020;12(1):76. doi: https://doi.org/10.3390/rs12010076.
[13] Ying H, Yong Z, Dunwei G. Multi-objective particle swarm optimization for feature selection with fuzzy cost. IEEE Trans Cybern 2021;51(2):874–88. doi: https://doi.org/10.1109/TCYB.2020.3015756.
[14] Chen K, Xue B, Zhang M, Zhou F. Correlation-guided updating strategy for feature selection in classification with surrogate-assisted particle swarm optimisation. IEEE Trans. Evol. Compu. 2021.
[15] Yong Z, Dun-wei G, Xiao-zhi G, Tian T, Xiao-yang S. Binary differential evolution with self-learning for multi-objective feature selection. Inf Sci 2020;507:67–85. doi: https://doi.org/10.1016/j.ins.2019.08.040.
[16] Xue Yu, Tang Y, Xu X, Liang J, Neri F. Multi-objective feature selection with missing data in classification. IEEE Trans Emerg Top Comput Intell 2022;6(2):355–64.
[17] Zhang J, Jia L, Menenti M, Hu G. Glacier facies mapping using a machine-learning algorithm: The Parlung Zangbo basin case study. Remote Sens 2019;11(4):452. doi: https://doi.org/10.3390/rs11040452.
[18] Khan AA, Jamil A, Hussain D, Taj M, Jabeen G, Malik MK. Machine-learning algorithms for mapping debris-covered glaciers: The Hunza basin case study. IEEE Access 2020;8:12725–34. doi: https://doi.org/10.1109/ACCESS.2020.2965768.
[19] Lu Y, Zhang Z, Huang D. Glacier mapping based on random forest algorithm: A case study over the Eastern Pamir. Water 2020; 12(11):323, Nov. 2020. doi: 10.3390/w12113231.
[20] Thanki D, Isrania D, Makwana A. Glacier mapping with object based image analysis method, case study of Mount Everest region. Jurnal Kejuruteraan/J Eng 2019;31(2):215–20. doi: https://doi.org/10.17576/jkukm-2019-31(2)-05.
[21] Racoviteanu AE, Nicholson L, Glasser NF, Miles E, Harrison S, Reynolds JM. Debris-covered glacier systems and associated glacial lake outburst flood hazards: Challenges and prospects. J Geol Soc 2022;179(3). doi: https://doi.org/10.1144/jgs2021-084.
[22] Agarwal V, Bolch T, Syed T, Pieczonka T, Strozzi T, Nagaich R. Area and mass changes of Siachen glacier (East Karakoram). J Glaciol 2017;63(237):148–63. doi: https://doi.org/10.1017/jog.2016.127.
[23] Zhou Y, Li Z, Li J. Slight glacier mass loss in the Karakoram region during the 1970s to 2000 revealed by KH-9 images and SRTM DEM. J Glaciol 2017;63(238):331–42. doi: https://doi.org/10.1017/jog.2016.142.
[24] Chavez S. Image-based atmospheric corrections-revisited and improved. Photogramm Eng Remote Sens 1996;62:1025–36.
[25] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans. Syst. Man Cybern Syst 1973;SMC-3(6):610–21.
[26] Song T, Duan Z, Liu J, Shi J, Yan F, Sheng S, et al. Comparison of four algorithms to retrieve land surface temperature using Landsat 8 satellite. Yaogan Xuebao/J Remote Sens 2015;19:451–64.
[27] Mölg N, Bolch T, Rastner P, Strozzi T, Paul F. A consistent glacier inventory for Karakoram and Pamir derived from Landsat data: Distribution of debris cover and mapping challenges. Earth Syst Sci Data 2018;10:1807–27. doi: https://doi.org/10.5194/essd-10-1807-2018.
[28] Bajracharya S (submitter), Guo W, Maharjan SB, Shrestha F (analyst(s)). GLIMS Glacier Database. Boulder, CO. National Snow and Ice Data Center 2014. doi :10.7265/N5V98602.
[29] Haireti A, Tateishi R, Alsaaideh B, S. Gharechelou S. Multi-criteria technique for mapping of debris-covered and clean-ice glaciers in the Shaksgam valley using Landsat TM and ASTER GDEM. J Mt Sci 2016; 13(4): 703–714. doi: 10.1007/s11629-015-3649-9.
[30] Thejas GS, Joshi SR, Iyengar SS, Sunitha NR, Badrinath P. Mini-batch normalized mutual information: A hybrid feature selection method. IEEE Access 2019;7:116875–85. doi: https://doi.org/10.1109/ACCESS.2019.2936346.
[31] Kayri M, Kayri I. The comparison of Gini and Twoing algorithms in terms of predictive ability and misclassification cost in data mining: An empirical study. Int J Compt Trends Technol 2015;27(1):21–30. doi: https://doi.org/10.14445/22312803/IJCTT-V27P105.
[32] Sharda S, Srivastava M. Optimized multiresolution segmentation for mapping glaciers. Indian J Ecol 2021;48(2):503–7.
[33] Chandra B, Paul Varghese P. Moving towards efficient decision tree construction. Inf Sci 2009;179(8):1059–69.
[34] Hall DK, Riggs GA, Salomonson VV. Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data. Remote Sens Environ 1995;54(2):127–40. doi: https://doi.org/10.1016/0034-4257(95)00137-P.
[35] Tielidze LG, Bolch T, Wheate RD, Kutuzov SS, Lavrentiev II, Zemp M. Supra-glacial debris cover changes in the greater Caucasus from 1986 to 2014. Cryosphere 2020;14(2):585–98. doi: https://doi.org/10.5194/tc-14-585-2020.
[36] Lv M, Guo H, Yan J, Wu K, Liu G, Lu X, et al. Distinguishing glaciers between surging and advancing by remote sensing: A case study in the Eastern Karakoram. Remote Sens 2020;12(14):2297. doi: https://doi.org/10.3390/rs12142297.
[37] Sharda S, Srivastava M. Object-based glacier surface change detection using multispectral satellite data. 2020 IEEE 7th Uttar Pradesh Sect. Int. Conf. Electr, Electron. Comput. Eng. (UPCON), Prayagraj, India, 2020. doi: 10.1109/UPCON50219.2020.9376524.