Subglacial water presence classification from polar radar data

Christopher M. Gifford a,*, Arvin Agah b,1

a National Security Technology Department, The Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA
b Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 666045, USA

Abstract

Ground and airborne radar depth-sounding of the Greenland and Antarctic ice sheets have been used for many years to remotely determine characteristics such as ice thickness, subglacial topography, and mass balance of large bodies of ice. Ice coring efforts have supported these radar data to provide ground truth for validation of the state (wet or frozen) of the interface between the bottom of the ice sheet and the underlying bedrock. Subglacial state governs the friction, flow speed, transport of material, and overall change of the ice sheet. In this paper, we utilize machine learning and classifier combination to model water presence from airborne polar radar data acquired on Greenland in 1999 and 2007. The underlying method results in radar independence, allowing model transfer from 1999 to 2007 radar data to produce water presence maps of the Greenland ice sheet with differing radars. We focus on how to construct a successful set of classifiers capable of high classification accuracy. Utilizing multiple machine learning algorithms is shown to be successful for this classification problem, achieving 86% classification accuracy in the best case. Several heuristics are presented for constructing teams of multiple classifiers for predicting subglacial water presence. The presented methodology could also be applied to radar data acquired over the Antarctic ice sheet.

1. Introduction

Remote sensing methods, such as radar and seismic/acoustic surveys, attempt to acquire data in order to infer properties of the subsurface from a remote location such as the surface, air, or space. A prominent example is the use of radar sensors to gather data from the polar ice sheets, specifically internal layers and the ice-bedrock interface, from the surface of an ice sheet. Other examples are satellite-based imagery and identification of land-cover and events, and the observation of light and its characteristics to infer properties of distant galaxies and stars.

Ground and airborne depth-sounding of the Greenland and Antarctic ice sheets have been used for many years to determine characteristics such as ice thickness, subglacial topography, and mass balance of large bodies of ice. Radar sounding of ice sheets is challenging due to the rough surface interface, various stages of melting both on top of and within the ice sheet, and spatial variation of ice thickness and bedrock topography. Processing the data, including the incorporation of knowledge about the sensing medium, is important for proper interpretation and dissemination of accurate data to the scientific community. For example, in Greenland, water is present over continuous finite distances which are smooth, but do not qualify as large lakes that are evident in Antarctica.

At the University of Kansas, the Center for Remote Sensing of Ice Sheets (CReSIS, 2009) performs polar research to gather data and model ice sheets to better understand global climate changes and the possible effects, including sea level impacts. We have designed, built, and utilized mobile robots to autonomously traverse polar terrain in Greenland and Antarctica and support radar remote sensing data acquisition (Gifford et al., 2009; Stansbury et al., 2004; Harmon et al., 2004; Akers et al., 2004, 2006a, 2006b). Airborne systems have been developed to offer large-scale studies of polar regions. One primary technology for this research is the development of sophisticated radars. CReSIS researchers have provided first views of the ice sheet bed in fast-flowing areas and certain internal ice sheet layers, and continue to provide detailed subsurface images and models describing their behavior and dynamics. CReSIS has a wide variety of radar data from various radar designs over many years. All of these data are available to the public, providing information such as latitude, longitude, radar travel times, bed echo intensity, and ice thickness for extended flight segments. This data repository can then be used for advanced machine learning, data mining, and modeling efforts for polar environments.

In this paper, we utilize ice-penetrating radar data collected in Greenland in May 1999 and September 2007 as part of a model-creation effort for subglacial water presence classification. Using radar
data from ice sheets, the goal is to learn a model of water presence or absence for each radar measurement. This enables the production of predicted water presence maps for the scientific community without needing to drill resource-intensive ice cores. Specifically, a detailed study of ensemble learning and decision combination has been conducted. Experimental results are presented, including successful ensemble compositions, individual learning algorithm contribution, team size, and team diversity. Finally, classification results as well as ensemble learning heuristics discovered as part of this investigation are discussed.

2. Background and related work

Machine learning techniques have seen limited application to ice sheet and polar subsurface data. Most efforts involve identification or tracing of specific layers that hold historical importance. Internal layers of ice sheets have been investigated to predict the depth and thickness of certain layers. For example, initial efforts to predict the depth and thickness of the Eemian Layer in the Greenland ice sheet utilized a Monte Carlo Inversion of the flow model to estimate unknown parameters constrained by the internal layers (Buchardt and Dahl-Jensen, 2007a, 2007b). One of these parameters is the basal melt rate, an important parameter for ice sheet models. Similar work has been done to classify the presence of bottom crevasses, and to estimate their height using radar data of the Ross Ice Shelf in Antarctica (Peters et al., 2007). A physical model was developed by studying the radar data and long echo tails that were found to be characteristic of bottom crevasses. This work largely involved studying the data to determine power reflection coefficients, and did not incorporate any regression or machine learning methods.

Learning classifiers have been recently employed for autonomous polar event detection via satellites. As part of NASA's New Millennium Program, the Autonomous Sciencecraft Experiment (ASE) has been used for detecting dynamic events—such as volcanic eruptions, floods, and cryospheric events—using onboard science algorithms. One such focus was on the detection of ice-based surface events (lakes freezing/thawing and sea ice breakup) (Castano et al., 2005, 2006). As part of that work, four pixel-based classifiers (manually constructed classifier, exhaustive threshold band ratio search classifier, decision tree, and SVM) were employed to generate models to recognize events using hyperspectral images. The SVM classifier, which has seen regular use onboard the EO-1 spacecraft, successfully classified sea ice breakup near Antarcctica which autonomously triggered a spacecraft reaction. Such a classifier can be used to identify high priority data and reduce the bandwidth for data communication back to Earth. In Srivastava and Stroeve (2003), preliminary results are presented for unsupervised discovery of geophysical processes (including snow, ice, and clouds) from a spaceborne instrument over Greenland. Their results showed that regional classifiers provide higher classification accuracy for polar regions, but still had difficulty differentiating between certain clouds and snow/ice. Others have taken a knowledge-engineering approach to build an intelligent satellite sea ice classifier by creating rules from sea ice experts (Soh and Tsatsoulis, 2000).

Efforts to classify subsurface layers, occasionally called facies in the literature, using ground penetrating radar (GPR) are abundant. Together with seismic surveying, GPR allows the remote sensing of subsurface properties to guide decisions of where to drill for oil or natural gas, as well as help explain observed surface changes. Radar reflection patterns can be used to distinguish echo returns in different regions, which lends well to machine learning methods to associate such patterns with distinct classes (Moysey et al., 2005). GPR signal interpretation is one of the more studied applications of subsurface classification, as interpretation of subsurface materials or layers requires coring or drilling to determine ground truth. Experts are required to manually study the materials to segment them into distinct rock or object classes. These data can then be used by classifiers to learn a model between GPR return signals and the excavated ground truth. As GPR has been used for many applications, machine learning can aid in automating these tasks by learning a model, interpolating between measurement sites, and characterizing the subsurface at other unknown locations.

From these works, it has been shown that using machine learning to create models utilizing polar radar data analysis is not only feasible, but also can provide high levels of accuracy while offering a significant increase in efficiency. In the following sections, we discuss the approach, using multiple learning algorithms, to further increase accuracy and efficiency for the application of subglacial water presence classification from radar data acquired on the Greenland ice sheet. This approach could also be applied to the Antarctic and other large bodies of ice, but is not explicitly discussed in this paper.

3. Radar, coring, and subglacial water

As part of NASA's Program for Arctic Regional Climate Assessment (PARCA) and others, the University of Kansas and CReSIS have collected ground and airborne radar data in Greenland which cover the majority of the continent's ice sheet. CReSIS now hosts these data for public use. Other major efforts, such as ice coring, have produced data that complement these radar data sets to study subglacial activity of the Greenland ice sheet.

Studying the extent/presence of subglacial water is important, as the state of the bed of the ice sheet (wet or frozen) provides information about flow, friction, and roughness of the underlying bedrock interface. The presence of water (wet state) means that the interface is likely slippery, or that heat has caused a downflow of water from upper or nearby portions of the ice sheet. Portions of the ice with larger amounts of water exhibit faster flow properties and dynamic changes of the ice in those regions due to lubrication at the bed. The lack of water means that the interface is frozen, likely exhibiting more friction and therefore little movement. A smooth bedrock interface lends more to sliding of the ice sheet along its surface, whereas a rough interface introduces more friction between the ice sheet and bedrock on which it rests. If water is present at the bed, then the interface between the ice and water will be smoother, compared to a mixed water and rock interface (which will exhibit the shape of the underlying rock). Transport of material, however, may cause this lower rock interface to change over time.

Ice coring offers an additional method to study, among many other things, the state of the ice-bedrock interface by manually examining the bottom of the cored ice column. Thus, ice cores represent the only ground truth available for large-scale ice sheet modeling and remote sensing validation. In Greenland, there are two primary drill sites which we focus on in this work, namely, GRIP and N-GRIP. The Greenland Ice Core Project (GRIP) drilled a core from 1989 to 1992 to a depth of 3029 m, located at 72.58 N, –37.63 W (NCDC, 2009). The subglacial state of this core was found to be frozen. The North GRIP (N-GRIP) drilled a core from 1999 to 2003 to a depth of 3085 m, located at 75.1 N, –42.32 W (NGRIP, 2009). The subglacial state of this core was found to be wet (high melt).

As a radar transmits energy down to and into the ice—typically from one or more transmitting antennas mounted on the wings of an aircraft—changes in dielectric properties cause the
signal energy to reflect and refract. Larger changes in dielectric properties (such as air to ice, and ice to rock) cause strong reflection intensities to be received by the radar's receiving antenna(s). Fig. 1 shows an example radar echogram over an ice sheet, illustrating the reflection of internal layers and the bedrock interface beneath the ice sheet. Each column of the echogram is a trace (right), representing sensor measurements with time (and therefore depth). Water is a comparatively strong reflector when surrounded by ice/rock and affects the signal return intensity, whether internal to or at the bottom of the ice sheet. Frozen ice against bedrock also represents a strong reflecting interface. Thus, other properties of the signal and the return envelope can be analyzed to determine whether water is present or not.

4. Combining decisions from multiple classifiers

Classifier combination is an area in machine learning that has offered advances in classification accuracy for complex data sets. It has been termed differently in the literature, namely, classifier fusion, mixture of experts, committees, ensembles, teams, pools, collective recognition, composite systems, etc. Several methods have been developed to combine predictions from multiple classifiers, the most popular being voting, boosting, bagging, and stacking.

The majority of efforts in combining classifiers have incorporated homogeneous learning algorithms. Fewer works have focused on combining heterogeneous learners, which is the focus of our study. The term “heterogeneous” is used in different contexts in the literature. Popular classifier combination approaches such as boosting and bagging involve homogeneous learning algorithms and manipulation of the training data set. Produced models are therefore homogeneous in representation; although the learners may output slightly different predictions. Use of different learning algorithms, and therefore heterogeneous model representations, is considered purely heterogeneous in the contexts of algorithm and model.

When developing a multi-classifier system, its members can be a mixture of weak (high error) and strong (low error) classifiers. Weak classifiers are typically simple to create, at the expense of their accuracy on complex data sets. Strong classifiers are typically time-consuming and expensive to create, as their parameters are fine-tuned for maximum performance. Furthermore, some classifiers perform better than others on the same data set due to their algorithmic nature. Combining weak/strong or homogeneous/heterogeneous classifiers offers the benefit of encompassing different levels of expertise and knowledge bases. Exploiting and studying these properties as advantages for classification is a driving force for multi-agent machine learning.

An important aspect related to difference in learners is their level of error correlation relative to one another, and as a whole. The more correlated (less disjoint) individual learners are, the less complementary they may be. If uncorrelated, they likely will misclassify different instances. Combining them better enables the system to correctly classify more instances. A significant improvement over a single classifier can only occur if the individual classifier theories are substantially different. It is desired to obtain a balance between high performance and complementarity in a team which combines decisions. If one learner does not predict correctly, the other learners should be able to do so. Diverse models are therefore more likely to produce error in different ways. Methods to accomplish this typically involve introducing diversity in terms of learning paradigms, feature subsets, or training sets (Tumer and Ghosh, 1996).

Although selection of the best individual classifier is easier and occasionally effective, combination techniques scale better to larger and more complex learning problems. Combining all classifiers in an ensemble can be improved upon by selecting for combination only those that perform significantly better than others, termed Selective Fusion (Tsoumakas et al., 2005; Kuncheva, 2002). Other efforts have studied the use of training multiple classifiers on different feature subsets prior to their combination (Chen et al., 1997). Focusing on differing and potentially overlapping feature subsets creates additional diversity, which could lead to an improved combined model. Techniques such as boosting and stacking can also be chained together to achieve high accuracy (Ebrahimpour, 2007). These works represent examples of how heterogeneity and diversity can lead to more robust, high-performance classifiers. Multi-classifier learning systems offer a medium for additional study on how knowledge from learners with different, potentially overlapping feature and data subsets can be combined. Interaction between the learners during the learning process may prove to increase learning efficiency, robustness, and accuracy.

5. Implementation and processing

The work presented here is based on a method that recovers radar reflection characteristics at the bed of the ice sheet that discriminate between frozen and wet subglacial properties for each trace of a radar survey (Oswald and Gogineni, 2008). The method is said to be most appropriate for Greenland, but a similar method could be applied to the Antarctic ice sheet. It is independent of the radar characteristics, as the two primary signal return attributes are radar independent if computed properly and normalized. Numerical and statistical analyses of radio-echo profiles and interface geometries, as well as ice and water physics, represent the basis of the method. Results were validated using known basal water presence at the GRIP and N-GRIP core sites. In Oswald and Gogineni (2008) more details of the method are provided, including related work on radio-echoing and depth-sounding of ice sheets.

A total of 13 strategic radar flights covering various portions of the Greenland ice sheet were flown throughout May 1999. These flight lines are shown in Fig. 2, with one line/color per flight. The radar data set used to develop the original method was collected from an aerial survey over Greenland on May 14, 1999.
Using the extracted basal water presence information by the published method for the 1999 ICARDS radar data, training and testing sets for classification were constructed for our machine learning study. This assumes regional similarity in water content and that the water presence has not drastically changed (prominently wet or frozen) at the GRIP and N-GRIP sites in the years from 1999 to 2007. Given that these sites are located near the Greenland ice divide, this assumption can be made without loss of generality.

This section presents the implementation details associated with processing the radar data for classification. The following sections discuss the implementation and results of a detailed multi-classifier machine learning study for the application of radar-independent model generation for polar subglacial research. The implementation incorporates additional information to that of the original method for more robust water presence classification and transfer of the concept to radar data acquired with new technology in the midst of global warming and climate change. Matlab (The MathWorks, 2006) was utilized for processing and manipulating the radar data into attributes for classification using the described method with slight modifications.

5.1. Smooth and filter data

Pre-processing of raw data consists of presumming (averaging) in hardware during acquisition, as well as offline coherent integration, range compression, frequency filtering, initial incoherent integration, azimuth compression, and waveform combination. This results in pre-processed data products, which are then passed through several cleaning stages before attributes can be computed using the received radar returns.

First, the data are incoherently integrated (using both magnitude and phase information for averaging) over an along-track distance of ~200 m for smoothing purposes. Next, individual traces are removed which exhibit the following "bad data" aspects:

- Latitude or longitude is erroneous (well away from traversed flight line, possibly due to poor GPS visibility during aircraft turns)
- Visually smeared surface or bed portions of preliminary-processed echogram
- Non-characteristic reflection magnitudes at bed, likely due to aircraft banking, nearby mountains or exposed rock, or large amounts of surface water
- Ice thickness is < 300 m (surface and bed reflections difficult to distinguish)
- Bed reflection is not visible
- Ice thickness or bottom depth could not be computed due to lack of visible bed reflection

Portions of the airborne survey that contain banks/tours of the aircraft may produce results that are not reliable or indicative of the true subglacial state. These portions are therefore removed prior to attribute calculation and processing.

5.2. Data calibration and boundary alignment

Most airborne ice sheet surveys begin on a known smooth, frozen path (runway) or portion of the ice sheet. This represents another known aspect of the ice sheet, to accompany ice cores, with which can be used for calibrating and bounding the intensity distributions received by the radar. These portions of the data are assumed to be indicative of typical frozen subglacial interfaces. The GRIP (frozen) and N-GRIP (wet) core sites provide reflection intensities that are known to range from frozen to wet. By focusing on these sources of ground truth, encompassing the

Specifications of the radar used to acquire these data, named Improved Coherent Antarctic and Arctic Radar Depth Sounder (ICARDS), is shown in Table 1. Uncertainty of the radar ice thickness measurements have been estimated to be ±10 m from comparisons to ice coring sites. We reproduced the results, with minor differences, of the published method, with minor modifications, on the same data set.

GPS technology allows the same flight lines to be flown for data acquisition over several years to study regional ice sheet changes due to climate and dynamics. In this work, we use the method’s radar independence to generate a subglacial water presence model for radar data acquired in 2007 from a new radar system, called the Multi-Channel Radar Depth Sounder (MCRDS), over an identical extended flight segment covering the GRIP and N-GRIP core sites. These flight lines (May 14, 1999 and September 17, 2007), which comprise the primary data set for our study, are shown in Fig. 3. Table 2 presents the specifications of the MCRDS radar system.

![Fig. 2. Flight lines for all airborne radar surveys flown throughout May 1999 over various regions of Greenland, with one line per flight. GRIP and N-GRIP core sites are also marked. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)](image-url)

using the described method with slight modifications.

![Table 1](image-url)

<table>
<thead>
<tr>
<th>Description</th>
<th>Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radar type</td>
<td>Pulse compression</td>
</tr>
<tr>
<td>Antennas</td>
<td>4-Element +/− dipole arrays</td>
</tr>
<tr>
<td>RF carrier frequency</td>
<td>150 MHz</td>
</tr>
<tr>
<td>RF Up-chirp bandwidth</td>
<td>17.00 MHz</td>
</tr>
<tr>
<td>Transmitted pulse width</td>
<td>1.6 ms</td>
</tr>
<tr>
<td>Compressed pulse width</td>
<td>60 ns</td>
</tr>
<tr>
<td>Peak transmit power</td>
<td>200 W</td>
</tr>
<tr>
<td>PRF</td>
<td>Up to 18.4 kHz</td>
</tr>
<tr>
<td>Coherent integrations</td>
<td>32–4096</td>
</tr>
<tr>
<td>Incoherent integrations</td>
<td>0–64,000</td>
</tr>
<tr>
<td>Baseband I &amp; Q bandwidth</td>
<td>8.5 MHz</td>
</tr>
<tr>
<td>A/D dynamic range</td>
<td>12-bit 72 dB</td>
</tr>
<tr>
<td>Sampling period</td>
<td>53.3 ns (18.75 MHz)</td>
</tr>
<tr>
<td>Sampling delay</td>
<td>0.080–300 ms</td>
</tr>
<tr>
<td>Pixel window in ice (range)</td>
<td>4.494 m</td>
</tr>
<tr>
<td>Trace spacing</td>
<td>~ 135 m</td>
</tr>
</tbody>
</table>

Table 1
ICARDS radar system specifications (Oswald and Gogineni, 2008).

First, the data are incoherently integrated (using both magnitude and phase information for averaging) over an along-track distance of ~200 m for smoothing purposes. Next, individual traces are removed which exhibit the following "bad data" aspects:

- Latitude or longitude is erroneous (well away from traversed flight line, possibly due to poor GPS visibility during aircraft turns)
- Visually smeared surface or bed portions of preliminary-processed echogram
- Non-characteristic reflection magnitudes at bed, likely due to aircraft banking, nearby mountains or exposed rock, or large amounts of surface water
- Ice thickness is < 300 m (surface and bed reflections difficult to distinguish)
- Bed reflection is not visible
- Ice thickness or bottom depth could not be computed due to lack of visible bed reflection

Portions of the airborne survey that contain banks/tours of the aircraft may produce results that are not reliable or indicative of the true subglacial state. These portions are therefore removed prior to attribute calculation and processing.

5.2. Data calibration and boundary alignment

Most airborne ice sheet surveys begin on a known smooth, frozen path (runway) or portion of the ice sheet. This represents another known aspect of the ice sheet, to accompany ice cores, with which can be used for calibrating and bounding the intensity distributions received by the radar. These portions of the data are assumed to be indicative of typical frozen subglacial interfaces. The GRIP (frozen) and N-GRIP (wet) core sites provide reflection intensities that are known to range from frozen to wet. By focusing on these sources of ground truth, encompassing the
The entire spectrum of subglacial water extent (fully wet to fully frozen), the method can be applied to the entire ice sheet. The observed lower intensity bound from these areas can be used as a reference. The upper intensity bound in the received intensities over flights represents wet interfaces with varying fractions of water content. Using the lower intensity bound, all extended flight segments have their intensity distributions shifted to exhibit matching lower bounds. This, if done correctly, removes the data’s dependence on radar settings and sensitivities, yielding radar independence.

During processing, received bed intensities are transformed from varying over 40 dB down to ~20 dB for more reliable basal reflection coefficient determination. Intense returns accompanied by high abruptness are indicative of the water presence. Large internal layer disturbances in the ice (ice stress and deformation) are also indicative of locations which likely contain water at the base. The original method verified this using the coherence index and echograms exhibiting such distortion (Oswald and Gogineni, 2008). The method underestimates the extent of water presence, producing a decision of frozen, where intensity is high but abruptness falls below threshold.

5.3. Radar signal attributes

Three primary radar data attributes are utilized by the method at each trace: relative signal intensity, signal abruptness index, and signal coherence index (Oswald and Gogineni, 2008). The first two are used to determine the binary result of wet or frozen, where relative signal intensity is the primary determinant. They employ coherence to further analyze and explain the results. These attributes are now discussed in more detail.

5.3.1. Relative intensity

This attribute represents the radar signal return intensity at the basal interface, including the first return and echo tail. The strength of the reflection intensity degrades with depth, as the signals attenuate and experience loss as they travel through solid ice and other media. Thus, the intensity variations can be modeled by an equation similar to the following to depth-adjust the received reflection due to rate of absorption in ice:

$$B = \frac{2.3 \times 3000}{H + 2000}$$

where $H$ represents the surface altitude, which governs intensity variation as a function of depth. The intensity after adjustment still requires recalibration for variations of the system, but is otherwise representative of the reflection coefficient at the bed of the ice. Pre-processed radar data go through several stages of processing and adjustment to arrive at the final relative intensity value of the bed reflection.

5.3.2. Abruptness index

This attribute provides information on whether the echo tail is beamwidth-limited. Abruptness constrains the angular spectrum of the received signal. The abruptness index can be computed as follows:

$$\text{Abruptness} = \frac{P_{\text{peak}}(X)}{P_{\text{agg}}(X)} = \frac{\max(P_{\text{window}}(X))}{\sum(P_{\text{window}}(X))}$$

The numerator in the calculation is the peak (max) basal reflection intensity ($P_{\text{window}}$) for a specific trace/position. The denominator is the aggregate (sum) basal reflection intensity, or the sum of return intensities over depth ($X$) bins in the basal...
return window \(P_{\text{window}}\). High values of abruptness are correlated with high aggregate intensity, but is not a clear discriminator of wet versus frozen on its own. Thus, high abruptness that is accompanied by intense returns is indicative of the presence of water. For the pulse length and sample rate used by the radar, the reflection abruptness is typically between 0.05 and 0.5. The method (Oswald and Gogineni, 2008) uses a value of 0.25 to identify locations that are significantly smooth, which is half of the maximum expected abruptness index.

5.3.3. Coherence index

This attribute provides information on whether the first Fresnel zone is determined by roughness or wavelength. Thus, coherence is another measure of smoothness at a smaller scale. Coherence can be calculated using the following equation:

\[
\text{Coherence} = \frac{\sum |\text{abs}(R_{\text{complex}}(D,X))|^2}{\sum |\text{abs}(R_{\text{complex}}(D,X))|^2}
\]

The numerator sums the complex echo signal \(R_{\text{complex}}(D,X)\) over the distance along-track \(D\) for coherent and incoherent integrations, which is evaluated for a slope that maximizes the coherent integral. The denominator is the signal power summed from the complex received signal and abruptness, independent of the received signal intensity. Coherent integrations are performed over a distance of \(\sim 200\) m. A coherence value of 0.3 or larger is indicated to surface facets that are flat to within 0.2 m over an integration distance of 200 m. As suggested in Oswald and Gogineni (2008), a coherence value of 0.3 and above, coupled with interface and surface slopes, produces a regression gradient that directly links smooth interface facets with physical characteristics of water.

Coherent integration is in the along-track direction, using only the magnitudes (no phase information) for averaging. Incoherent integration is also in the along-track direction, but uses both magnitude and phase information for averaging. Received intensities are incoherently averaged over a distance of \(\sim 200\) m to achieve smoothing and reduce signal intensity variation. For smooth interfaces, the response is more impulse-like; whereas slightly rough interfaces exhibit more intensity variation; and rough interfaces produce intensity variations more evenly distributed. The intensity distribution is largely caused by the scatter of energy as a function of interface roughness. Variance of the received signal is further reduced by aggregating the basal reflection envelope over all beam pattern directions (Oswald and Gogineni, 2008).

5.4. Computing data attributes

Prior to computing data attributes, each trace has the bed reflection (peak intensity depth from expected bed region of A-scope, where visible) selected by a radar expert, assisted by surface- and bed-picking software. Fig. 4 demonstrates this process using the custom software, showing echograms surface returns fully picked and the bed reflections partially picked. The surface can be easily selected, as it typically represents the first and largest reflection intensity spike in the trace's return profile. Using the corresponding radar echo time, peak depth and ice thickness can be calculated.

A five-sample window is created around the peak bed reflection for each trace, containing two samples above and below the determined bedrock reflection depth. Abruptness is computed using the peak reflection and the sum of the bed reflection window intensities (the depth bins). The received reflections are depth-adjusted, as previously described, using the ice thickness at each trace. Peak reflection intensity is converted to a relative intensity value in dB by dividing the peak at each trace by the maximum peak intensity in the data's basal reflection window.

An important aspect of this method is the alignment of the lower boundary of the intensity distributions. The upper boundary corresponds to the wet interface, while the lower boundary corresponds to the rock/frozen interface. By studying the ground-truth GRIP (frozen) and N-GRIP (wet) coring sites, this lower/ frozen reflection intensity boundary can be found. For radar data acquired over these core sites in 1999, this lower bound was found to be \(\approx 31.37\) dB. Each flight's lower boundary is aligned using this reference value to perform the final step of removing the effects of radar settings.

Fig. 5 plots the relative intensity distribution over the May 14, 1999 flight segment containing wet and frozen bed characteristics, illustrating a bimodal distribution (one hypothetical Gaussian distribution for the wet intensities and another for the frozen intensities). Some portions of the 20–30 dB variation in intensity overlap at lower probabilities. The overlapping area is where it can be distinguished between wet (higher relative reflection intensities) and frozen (lower relative reflection intensities) areas by computing a threshold which separates the bimodal distribution into wet and frozen portions. Observed intensities showed a separation of \(\approx 13\) dB between the mean of each Gaussian mode with high confidence (Oswald and Gogineni, 2008). Once the thresholds and lower bound of the intensity distribution are found, the method can be applied to radar data acquired over the entire Greenland ice sheet to recover the extent of basal water presence. Finally, coherence is calculated for each trace.

Fig. 6 presents the attributes, thresholds, and binary basal water presence classification for the extended flight segment 42–70 from the May 14, 1999 flight. Along with the echogram for this segment, the bedrock reflection window is shown, on which relative intensity calculations are based. The empirically-found thresholds of \(-15\) dB (relative intensity) and 0.27 (abruptness) dictate water presence for each radar trace. Coherence values above 0.3 represent smoother surfaces, which are utilized as additional signal information for classification. The referenced study (Oswald and Gogineni, 2008) uses a value of 0.25 to identify locations that are significantly smooth. We found an abruptness value of 0.27 (slightly more than half of the maximum abruptness index value) more closely matched our
ground truth data, provided that the referenced method did not explicitly specify the relative intensity used for experimentation. Our empirical results suggest a relative intensity value of $-15$ dB, coupled with 0.27 for abruptness and 0.3 for coherence, provided a best fit to our ground truth water presence data, as well as alignment with results obtained by others (Oswald and Gogineni, 2008).

6. 1999 ICARDS radar data set

Each trace is marked as being wet or frozen using the calculated abruptness and relative intensity for each trace. A trace, $t$, is determined to contain water at the bedrock if:

$$\text{RelIntensity}(t) \geq -15.0 \quad \text{and} \quad \text{Abruptness}(t) \geq 0.27$$
The binary (wet/frozen) 1999 ICARDS data set is then generated by saving the following attributes for each trace:

- Latitude/longitude
- Relative intensity (dB)
- Abruptness ([0,1])
- Coherence ([0,1])
- Ice thickness (m)
- Subglacial state: wet (1), frozen (0)

Ice thickness is included to allow the learning algorithms to take into account depth when fitting a model to the data for classification. All three primary attributes are used, plus others, for learning a model connecting the radar data and subglacial water presence. The original method’s paper reports results using only relative intensity and abruptness attributes, with some analysis using coherence (Oswald and Gogineni, 2008).

Each trace of the radar survey represents an instance which can be used for training or testing a classifier. The above information represents the attributes for each instance. Finally, the method’s output represents the instance’s class, used as the target for training a classifier or for computing classification accuracy during testing. This represents a binary classification problem. The data set’s statistics are shown in Table 3.

### Table 3
1999 ICARDS binary data set statistics.

<table>
<thead>
<tr>
<th>Data set traces (instances)</th>
<th>Total</th>
<th>Wet</th>
<th>Frozen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>-34.3455</td>
<td>0.0455</td>
<td>0.0007</td>
</tr>
<tr>
<td>Abruptness</td>
<td>0.16217</td>
<td>0.4115</td>
<td>0.8331</td>
</tr>
<tr>
<td>Coherence</td>
<td>0.33872</td>
<td>0.4115</td>
<td>0.8331</td>
</tr>
<tr>
<td>Thickness</td>
<td>-18.0785</td>
<td>0.2332</td>
<td>0.1073</td>
</tr>
</tbody>
</table>

Using the described method, water presence has been classified for the entire May 1999 Greenland radar survey. The method classifies that, out of a total of 7995 traces for this segment, 1565 are wet and 6430 are frozen. The GRIP and N-GRIP sites are classified as frozen and wet, respectively, and water tends to be present in continuous regions. Fig. 7 shows the results spatially. Icons marking water presence over flight lines are sized for visualization purposes, and should not be interpreted as containing water beneath the entire icon.

#### 7. 2007 MCRDS radar data set

As the data and method were studied and developed in detail, the May 1999 ICARDS radar data set represents ground truth for further studies. This portion of the 1999 data set encompasses some of the limited pure ground truth available in Greenland. This data set is used in our work to train teams of classifiers to learn a model between radar surveys acquired with different radars over the Greenland ice sheet.

The corresponding flight segment from the 2007 MCRDS Greenland radar survey was flown on September 17, 2007. In particular, the D-K data segments form a flight path that overlaps the same flight path flown over GRIP and N-GRIP in 1999. We can thus use the method’s output from 1999 data at those locations to provide the target class (wet or frozen) for training using the 2007 data.

The 2007 MCRDS radar data are processed using the same method as the 1999 data, with the following minor differences:

- Use of seven traces on each side of the current trace (15 total traces) for incoherent integration and computing coherence (as opposed to 1 on each side, 3 total traces, for the 1999 data).
- Differences between the oscillators used in the 2007 radar, which affects the coherence attribute. This is due to improved technology.
- Alignment of the lower boundary of the reflection distributions of the 2007 data is performed using reflection intensity bounds from the MCRDS data.

Fig. 8 plots the relative intensity distribution over the September 17 flight segment from the 2007 Greenland MCRDS radar survey. Again, a bimodal distribution is observed: one hypothetical Gaussian distribution for the wet intensities, and another for the frozen intensities. However, compared to the 1999 ICARDS relative intensity distribution, it is more difficult to distinguish between wet and frozen basal states, as the distribution exhibits more overlap.

The 2007 MCRDS data set is created using attributes computed from the MCRDS radar data and the geographically closest location’s class (wet or frozen) from the 1999 ICARDS model. Fig. 9 shows the transferred water presence (marked with icons) for training and testing purposes. This assumes regional similarity in subglacial state, and the general stability of this state over the time difference between data acquisition between these systems. Thus, the ground truth water presence output from the described method is used to label the 2007 MCRDS data for each trace’s water presence. The reason that this is possible, even with heterogeneous radars, is that attributes of the 2007 processed data are normalized independently, and independent of the 1999 data. Even though the reflection intensity, abruptness, and coherence magnitudes are different from radar to radar and with different

![Fig. 7. Water presence for the extended segment of the May 14, 1999 flight surrounding GRIP and N-GRIP sites. Icons mark the presence of water.](image-url)
radar settings, a common classification target and normalization facilitate model transfer from one year’s survey to another. The 2007 data set comprised the same attributes as the 1999 ICARDS data set. Fig. 10 presents the attributes and binary basal water presence classification for the extended flight segment D-K from the September 17, 2007 flight. Along with the echogram for this segment, the bedrock reflection window is shown, on which relative intensity calculations are based. Training and testing sets are created by randomly splitting the data into two-third training and one-third testing portions. The statistics of this data set and its training and testing portions are shown in Table 4. Out of a total of 45,477 traces, 7819 are wet and 37,658 are frozen.

The number of frozen training instances greatly outnumbers that of wet training instances (nearly fivefold), creating a highly unbalanced training problem. As the number of ground truth sites were limited (i.e., where drilling had occurred from surface to bedrock to determine wetness, coupled with multiple radar passes acquired over time), we approached the training problem from a site-based perspective. Further, radar data processing and experimentation suggested that there were varying levels of “frozen” from radar signature and basal wetness/melt perspectives. As such, we aimed to capture as much of the variation as possible with a larger amount of frozen bedrock examples from various tracks, while including all wet bedrock instances for model creation. Our primary concern was identifying water presence with confidence. Therefore, the levels of “frozen” were accumulated into a single class.

In the process of constructing the classifiers to approach this problem, experiments were conducted with other training set compositions which employed sampling techniques to reduce the size of the training set (e.g., number of total training instances, due to the resolution of the employed radars) and attempt to avoid overfitting. It was found that training classifiers with more radar instances, specifically for those traces labeled as frozen in the ground truth data, resulted in a more robust and accurate classifier set. This became apparent when testing the classifiers with radar data that was spatially (i.e., regionally) similar or exhibited characteristics similar to the sites of interest (i.e., GRIP and N-GRIP). Although the data set was unbalanced, little adverse overfitting was observed. As ground truth sites are drilled and more data becomes available via drill site analysis and radar transects, the need to sample for training purposes will be explored.

The 2007 data set is provided to ensembles of classification algorithms to create a combined model of the radar data against the determined subglacial state for all traces. Classification results offer probability estimates for each trace’s inferred state (wet, frozen). Probability estimates from multiple learning algorithms, using different model fitting techniques, provide further information with which to determine the extent of the ice sheet’s subglacial water presence.

8. Research methodology

8.1. Multi-Classifier Collaborative Learning Architecture

The WEKA machine learning suite (Witten and Frank, 2005) is utilized as a base for the implementation of the proposed Multi-Classifier Collaborative Learning Architecture (Gifford, 2009). The primary uses of WEKA in this architecture are to prepare the data for experimentation and provide implementations of various classification algorithms. Additional Java and script/batch file implementations act as a wrapper for machine learning experiments involving single or multiple learners (i.e., for teams of any size), homogeneous or heterogeneous team composition, independent or collaborative learning (with the ability to vary
the number of collaboration events during learning), and combining the decisions of the learners using a variety of accuracy- and vote-based combination techniques. This offers a robust and flexible architecture for machine learning studies involving multiple classifiers.

The Multi-Classifier Collaborative Learning Architecture was utilized to study aspects of team learning for the subglacial water presence classification data set. By default, each learner is provided a randomized version (different orders of instances for each learner) of the full training data set. Learning takes place by training each classifier using the training data set. At this time, each learner tests its model on the data on which it was trained. Once training is complete, the testing data set is passed through each learner’s model and the corresponding class probabilities (predictions) are recorded. The final individual testing predictions are then used for combining decisions from multiple learners. This acts as the final collaboration step, which fuses the knowledge from multiple learners to a single team classifier via accuracy- and vote-based mechanisms.

Decision combination utilizes each learner’s classification for each testing instance to arrive at a single team classification per combination method. A learner’s prediction consists of a probability that the testing instance belongs to each of the possible classes. The highest probability represents the predicted class for each classifier. The following vote-based combination methods are used for evaluation. Selective variations of all methods listed below, with the exception of Select Best, were also examined. Selective variations for each method only take into account learners that have a testing accuracy above the mean team testing accuracy. Ties are broken by using the class selected by the learner with the highest accuracy on the testing data, unless otherwise stated. The combination method resulting in the best classification accuracy is selected, reflecting overall team performance.

Max: Selects the class with the absolute maximum probability value out of all learners.

Average: Averages all predictions for each class from all learners and selects the class corresponding to the highest average value.

Multiply: Multiplies all prediction values for each class from all learners and selects the class corresponding to the highest resulting value.

Table 4
2007 MCRDS binary data set statistics.

| Training set | Total 30 318 | 5213 | 25 105 |
| Testing set | Total 15 159 | 2606 | 12 553 |
| Intensity | Abruptness | Coherence | Thickness |
| Min | – 31.3996 | 0.1589 | – 3.9975 | 2675.1224 |
| Max | – 2.47822 | 0.3045 | 0.99843 | 3306.2202 |
| Avg | – 24.0789 | 0.2054 | – 3.7301 | 3062.4608 |

Fig. 10. Attributes and water presence classifications for each radar trace of the D-K extended flight segment from September 17, 2007. GRIP and N-GRIP core locations are shown for reference.
Majority Vote: Counts the number of votes (highest class probability) for each class from all learners. The class with the most votes wins. Ties are broken by using the first occurrence of the highest sum of testing accuracies from voters of the class receiving the most votes.

Weighted Vote: The class vote from each learner is multiplied by its accuracy on the training data (as a percentage). All votes are then summed, and the class with the highest vote tally wins. This essentially represents the learner with the highest accuracy on the training data having the most influence in which class is selected. Ties are broken by using the first occurrence of the maximum vote value.

Select Best: The learner with the highest accuracy on the training data is used for all final predictions.

8.2. Experimental setup

A tournament-style process with rounds based on team size is employed to perform a comprehensive study on all aspects of ensemble learning for this application. Team configurations exhibiting the highest testing accuracy move on to the next round, where larger teams are constructed from the best teams from the previous round.

The first step for a specific data set is to coarsely test a variety of individual machine learning algorithms, each with a few different settings and initialization seeds, to determine which algorithms perform well in general for the data set. The best settings and corresponding 10-fold cross-validation testing accuracy for each algorithm are recorded, and the top five algorithms are selected to advance to the next round, consisting of teams of size two.

All pairs of these top five algorithms with their best individual settings are then run. This inherently includes homogeneous and heterogeneous team compositions. Similarly, the top teams of size two are selected based on combined testing accuracy. The same process takes place for teams of size four, where the top teams of size four are selected to advance to the final round of size eight teams. Once the round of size eight teams is complete, all results are compiled and a full study can take place.

Based on our experience with multi-classifier systems and applying them to other real-world data sets, we have found that teams larger than size eight typically do not offer a significant accuracy versus cost tradeoff. What we found to not offer a statistically significant improvement in performance. Average accuracy as team size increased exhibited a plateau at and beyond teams of size eight for this classification problem.

The goal of evaluating these aspects through detailed team studies is to study whether teams are better than an individual, combining decisions from multiple learners is useful, and that heterogeneous learning algorithms leads to higher task performance/accuracy. The following sections present experimental results and discuss general aspects of team learning discovered as part of this research.

9. Experimental results

This section presents the experimental results on various aspects of team learning from the subglacial water presence classification study. The notation used in the subsequent figures and analyses are abbreviations of learning algorithms. The abbreviations for the 11 learning algorithms are Naive Bayes (NB) (John and Langley, 1995), Decision Tree (DT) (Quinlan, 1993), Instance-Based KNN (IBK) (Aha et al., 1991), Neural Network (NN), Logistic Regression (LGR) (Le Cessie and van Houwelingen, 1992), Radial Basis Function Network (RBF), K* (KST) (Cleary and Trigg, 1995), Decision Table (DTB) (Kohavi, 1995), RIPPER Rule Learner (JRP) (Cohen, 1995), PART Rule Learner (PRT) (Frank and Witten, 1998), and Random Forest (RFT) (Breiman, 2001). The reader is referred to the referenced papers for the algorithms for detail on their underlying structure and theory. When listing team compositions, the number of learners for each learning algorithm of the team is listed, separated by “+”. For example, “3ibk+2nn+2dt+1lgr” represents a team of size eight composed of three IBK, two NN, two DT, and one LGR classifiers.

A total of 73 learning experiments were performed, including the variation of team size and composition. Experiments are broken down as follows:

- 11 size one, 19 size two, 24 size four, and 19 size eight
- 26 homogeneous and 47 heterogeneous teams

9.1. Team rankings

First, team compositions that are generally successful were investigated. Table 5 shows the best and worst 10 teams overall, based on testing accuracy, accompanied by the associated team sizes. Teams of size eight produce the highest team classification accuracy. In general, larger team sizes produce higher testing accuracies for this data set, while smaller teams produce the lowest testing accuracies. Certain pairings of learning algorithms perform very poorly in comparison to the majority of the teams as well. Increasing the number of learners in the team can help stabilize team accuracy. Team composition makes a difference in team accuracy, as each learning algorithm models the data in a different manner. Combining these different models in different ways can increase accuracy.

Two primary trends are apparent when analyzing the best performing teams: (1) heterogeneous teams consistently produce high team classification accuracy, and (2) larger teams (d Dominantly size eight in the top 10) consistently perform better.
9.3. Single versus multiple classifiers

A study of ensemble learning inherently involves the question, “How large should the ensemble be?” Specifically, it is desirable to study how scaling the team size affects different dynamics (successful team composition, combined testing accuracy, etc.). The objective is to study what can be gained by using multiple learning algorithms, as opposed to a single learning algorithm. The following table sheds more light on these aspects, and is intended to supplement the results presented as part of other studies in this paper.

Table 6 lists the five most successful and least successful teams as a function of team size. This table shows the direct result of using a tournament-style experimentation setup, as the best individual learners are heavily used in all successful teams of larger sizes. A tradeoff does exist between team size and classification accuracy. Team size inherently involves a difference in training and testing time. Some of the learning algorithms require more resources and time to perform their modeling effort. For example, decision trees do not require the memory space and time that neural networks do. The K* algorithm is the most memory-intensive algorithm utilized in the experiments. Thus, if similar accuracy can be achieved with a smaller team size and the application permits, it may not be efficient to utilize larger, more resource-intensive teams for a minimal amount of accuracy gain.

There exists a coupling between team size and combined testing accuracy: testing accuracy increases with team size. Again, evidence that RFT is the best individual classifier is apparent, also representing a strong base classifier with which to construct an ensemble. This suggests, for this application, that a strong base classifier is more important than team diversity. A heuristic which follows from these results is to construct a base homogeneous

Smaller team size is indicated as a prominent trend for the worst performing teams. The majority of poor-performing teams are combinations of poorly performing individual classifiers, or homogeneous teams of the same poor individual classifier on this data set. There is a clear separation between those learning algorithms that offer high accuracy, and those that do not.

It can additionally be seen that multiple teams can produce the same testing accuracy. For example, a base team of six RFT classifiers is made better by adding one or two other classifiers. DT, KST, and DTB can be added to offer additional viewpoints and decisions, all producing the same combined testing accuracy. In this case, these classifiers are required to achieve higher accuracy than a team of only six RFT classifiers. The worst teams perform only 3% less than the best teams, which are largely composed of individual classifiers. Thus, it can be concluded that larger team sizes, coupled with heterogeneity, provide a small improvement for this binary classification problem.

9.2. Team diversity

Team diversity, or the composition of a team in terms of multiple heterogeneous learning algorithms, is central to a study of ensemble learning dynamics. As different learning algorithms model the data and error differently, combining them can be beneficial for increasing classification accuracy. Teams composed of the same learning algorithm may not experience these advantages, especially when collaboration is involved. Here, we specifically focus on comparing results from homogeneous and heterogeneous teams. When discussing a team's diversity level, it represents the number of unique learning algorithms in that team.

One method for analyzing team diversity is to observe its effect on testing accuracy as a function of the diversity level, as shown in Fig. 11. No team existed which contained five, six, or seven unique learning algorithms, and the testing accuracies are assumed linear between diversity level four and eight. First, the tournament-style experimentation process produced many teams of diversity levels 1, 2, and 3. Teams see a steady increase in testing accuracy coupled with an increase in diversity up to a diversity level of three, after which accuracy levels off. Thus, a team consisting of three heterogeneous learning algorithms is desirable in this application, as it maximizes testing accuracy.

Further, heterogeneous teams (i.e., those containing more than one unique learning algorithm) perform 3% better on average compared to homogeneous teams. Homogeneous team compositions perform best at larger team sizes, and worst as individual classifiers. Increasing team diversity has proven necessary to offer an increase in testing accuracy. These results again stress the importance of team composition as a driving factor for higher classification accuracy.

<table>
<thead>
<tr>
<th>Best 10 teams</th>
<th>Worst 10 teams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Team</strong></td>
<td><strong>Size</strong></td>
</tr>
<tr>
<td>7rft + 1 dt</td>
<td>8</td>
</tr>
<tr>
<td>7rft + 1 db</td>
<td>8</td>
</tr>
<tr>
<td>7rft + 1 kst</td>
<td>8</td>
</tr>
<tr>
<td>8rft</td>
<td>8</td>
</tr>
<tr>
<td>6rft + 2 nn</td>
<td>8</td>
</tr>
<tr>
<td>5rft + 2 nn + 1 db</td>
<td>8</td>
</tr>
<tr>
<td>5rft + 2 nn + 1 dt</td>
<td>8</td>
</tr>
<tr>
<td>6rft + 1 dt + 1 kst</td>
<td>8</td>
</tr>
<tr>
<td>6rft + 1 db + 1 dt</td>
<td>8</td>
</tr>
<tr>
<td>6rft + 1 db + 1 kst</td>
<td>8</td>
</tr>
</tbody>
</table>

| Table 6 Overall best and worst 10 teams, ranked by combined testing accuracy. |
|------------------------|------------------------|
| **Best teams**         | **Worst teams**        |
| **Team**               | **Size**               | **Accuracy** | **Team**               | **Size**               | **Accuracy** |
| 7rft + 1 dt            | 8                       | 0.8572       | 1ibk                   | 1                       | 0.8264       |
| 7rft + 1 db            | 8                       | 0.8567       | 1nb                    | 1                       | 0.8265       |
| 7rft + 1 kst           | 8                       | 0.8567       | 1rft                   | 1                       | 0.8270       |
| 8rft                   | 8                       | 0.8565       | 1rft                   | 1                       | 0.8281       |
| 6rft + 2 nn            | 8                       | 0.8563       | 1rpt                   | 1                       | 0.8315       |
| 5rft + 2 nn + 1 db     | 8                       | 0.8562       | 2db                    | 2                       | 0.8338       |
| 5rft + 2 nn + 1 dt     | 8                       | 0.8550       | 4db                    | 4                       | 0.8339       |
| 6rft + 1 dt + 1 kst    | 8                       | 0.8559       | 1rpt                   | 1                       | 0.8340       |
| 6rft + 1 db + 1 dt     | 8                       | 0.8559       | 2rpt                   | 2                       | 0.8341       |
| 6rft + 1 db + 1 kst    | 8                       | 0.8559       | 1db                    | 1                       | 0.8342       |

| Table 6 Overall best and worst 10 teams, ranked by combined testing accuracy. |
|------------------------|------------------------|
| **Best teams**         | **Worst teams**        |
| **Team**               | **Size**               | **Accuracy** | **Team**               | **Size**               | **Accuracy** |
| 7rft + 1 dt            | 8                       | 0.8572       | 1ibk                   | 1                       | 0.8264       |
| 7rft + 1 db            | 8                       | 0.8567       | 1nb                    | 1                       | 0.8265       |
| 7rft + 1 kst           | 8                       | 0.8567       | 1rft                   | 1                       | 0.8270       |
| 8rft                   | 8                       | 0.8565       | 1rft                   | 1                       | 0.8281       |
| 6rft + 2 nn            | 8                       | 0.8563       | 1rpt                   | 1                       | 0.8315       |
| 5rft + 2 nn + 1 db     | 8                       | 0.8562       | 2db                    | 2                       | 0.8338       |
| 5rft + 2 nn + 1 dt     | 8                       | 0.8550       | 4db                    | 4                       | 0.8339       |
| 6rft + 1 dt + 1 kst    | 8                       | 0.8559       | 1rpt                   | 1                       | 0.8340       |
| 6rft + 1 db + 1 dt     | 8                       | 0.8559       | 2rpt                   | 2                       | 0.8341       |
| 6rft + 1 db + 1 kst    | 8                       | 0.8559       | 1db                    | 1                       | 0.8342       |
team of a strong classifier, and add one or two other potentially complementary classifiers. For large teams, this could translate to a heterogeneous mixture of homogeneous teams.

Conversely, the least successful teams over all team sizes generally do not follow this heuristic. Some of the poor teams exhibit diversity equality (team size divided by diversity level of 1 to 2). This was unexpected, as the theory of combining multiple heterogeneous learners as part of a highly diverse team of classifiers suggests higher accuracies and increased learning stability. In general, however, homogeneous teams offer comparable, but slightly lower, combined testing accuracy. Teams of size five to seven were not tested as part of this study.

9.4. Decision combination method

Decision combination methods are the means for combining classifications from multiple learning algorithms to formulate the ensemble’s collective decision for an instance requiring classification. In the following table, each count for a combination method represents an experiment which resulted in that combination method providing the highest testing accuracy. Ties (in terms of combined decision testing accuracy from all learners) count as a win for each method. The reported win percentages have been normalized, as there are varying numbers of teams of different sizes and more heterogeneous teams, due to heterogeneous teams generally offering higher classification accuracy in the tournament-style experimentation process.

Table 7 presents the most successful five decision combination methods over all experiments for this data set. In general, team size dictates the combination method to choose. Multiply, Average, and their Selective variations are successful methods for these experiments. It was found that Max performs best for size two, Selective Average for size four, and Selective Multiply for size eight teams. Select Best, Majority Vote, and Weighted Vote performed comparatively poor for all team sizes. Max, Average, and Multiply are best for homogeneous teams, while Selective methods perform best for heterogeneous teams. These results suggest that the Multiply method is superior for this classification problem, followed closely by Average and their Selective variations.

9.5. Algorithm contribution

The contribution and success of individual machine learning algorithms were investigated. By analyzing a learning algorithm’s contribution to success, learning algorithms that are generally successful for this application, and those which also perform well together, can be extracted. If teams that contain a certain learning algorithm consistently perform better than others not containing that learning algorithm, then that algorithm can be considered valuable (or a major contributor) to a team’s success.

This is measured in two primary ways. First, as shown in Fig. 12, average testing accuracy is compared for each learning algorithm over all teams in which that algorithm participated. On average, RFT, IBK, KST, and NN perform the best. Teams containing NB, DTB, JRP, and RBF learners are less stable, as they can produce very poor classification accuracy when included in a team. On the other hand, learning algorithms such as RFT, IBK, and KST are stable learning algorithm choices, as their minimum team accuracies are still comparably high. Combining these algorithms is encouraged, to produce a team that has a slightly heterogeneous mixture of strong, complementary classifiers.

Second, as shown in Fig. 13, the number of total teams which each learning algorithm participated in can be compared. This figure shows individual learning algorithm 10-fold cross-validated accuracies, and learning algorithm participation over all team learning experiments as a percentage. The byproduct of the tournament-style experimental setup is the use of the best performing algorithms in more teams. It is also evident that the RFT, IBK, and KST classifiers are used in 20% more teams than the next best set of classifiers.

Table 7

<table>
<thead>
<tr>
<th>Combination method</th>
<th>Win count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiply</td>
<td>29</td>
</tr>
<tr>
<td>Selective multiply</td>
<td>25</td>
</tr>
<tr>
<td>Max</td>
<td>24</td>
</tr>
<tr>
<td>Average</td>
<td>19</td>
</tr>
<tr>
<td>Selective average</td>
<td>19</td>
</tr>
</tbody>
</table>

Fig. 12. Minimum, maximum, and average testing accuracies over all teams containing one or more of each learning algorithm.
This is a result of combining teams of size two to form teams of size four, and combining teams of size four to form teams of size eight. NB, DTB, JRP, and RBF classifiers were utilized least overall.


Using the established thresholds for the 1999 radar data and the classification model created for the 2007 radar data set, the accuracy of the 2007 water presence to the accepted 1999 ground truth results surrounding the GRIP and N-GRIP core sites can be compared. Fig. 14 presents the comparison of water presence classification for the geographically closest 1999 and 2007 flight segments. The model corresponding to the best team of classifiers (7RFT + 1DT, 85.72% combined testing accuracy) was utilized to produce 2007 water presence.

The ice sheet’s bedrock interface topography is very similar between the two data sets, except where the 2007 flight line gets further away. The 2007 survey has much higher data resolution compared to the 1999 survey, and the icon size used to indicate water presence is not to scale. The GRIP and N-GRIP sites are accurately classified as dominantly frozen and wet, respectively. The majority of misclassified radar traces are in predominantly...
wet areas, occasionally being incorrectly classified over steeper bedrock slopes. The northernmost portion of the 2007 flight segment is also classified as containing several areas of subglacial water presence, which were not present at a similar scale in the 1999 data.

Machine learning was able to create an accurate model of subglacial water presence, utilizing 1999 ground truth data as a means to transfer water presence to 2007 radar data. These results demonstrate that pattern recognition techniques can be important tools for analyzing radar and other data from the Greenland ice sheet. By integrating more attributes about the ice sheet from various sensors and existing models, water presence indicators can be discovered and employed in the creation of maps showing the inferred subglacial state.

10. Conclusion

The work discussed in this paper represents a unique classification problem, attempting to associate remotely sensed radar data properties to subglacial water presence and transfer the model to different radar systems. To our knowledge, this is the first attempt to classify water presence from radar data acquired over the polar regions using multiple machine learning algorithms. Knowledge of ice dynamics, ground truth core data, and attribute extraction from airborne radar returns have allowed us to learn a model of subglacial water presence to facilitate in the creation of water presence maps of the entire Greenland ice sheet. Using methods to achieve radar independence, attributes were extracted to create a binary (wet or frozen) data set from radar data acquired in 1999 and 2007. It is planned to extend this concept to Antarctic radar data as well, such that water presence maps can be generated and monitored from new radar data as they become available.

Assumptions made as part of this classification problem and general study could, in part, contribute to the similar accuracy of most classifier ensembles. Specifically, the radar data were acquired over the ice sheet eight years apart. With global warming and climate change being of increased focus and research in recent years, this time difference could have allowed for small changes on top of, within, and at the base of the ice. These changes could, in turn, affect the radar data and therefore the method utilized to achieve radar independence in a manner which affects classification. For instance, temporal temperature, snow accumulation, and/or melting alter the ice sheet and therefore the data recorded by the sensor. Radar settings and hardware differences may also contribute to small-scale changes in the radar data and measured intensity at the subglacial interface. However, at the locations where these data were acquired, subglacial water presence is assumed to be relatively static over the time difference between data sets.

Using the proposed Multi-Classifier Collaborative Learning Architecture, ~86% classification accuracy was obtained. Several heuristics were discussed for constructing teams of multiple classifiers for this application. A tradeoff between team size and team composition, including the time required to complete the learning process, also became apparent. The Random Forest (RFT) classifier was the best individual classifier, as well as a necessary base for increased classification accuracy in heterogeneous teams. Strong teams were composed of a base homogeneous team of RFT classifiers accompanied by one or two other classifiers which were complementary. Larger teams proved to be well suited for this data set, where heterogeneous teams with low diversity performed best on average. Heterogeneous algorithms offer the advantage to integrate different error models. Contrary to our expectations, equally diverse teams (team size divided by diversity level of 1 to 2) were not among the best for this data set.

There are many facets of team learning that require further investigation. Some real-world applications are restricted by limitations of computing power, working memory, and decision time. Additional future work could be performed which focuses on groups of classifiers that have various limitations on memory, time, and processing. General heuristics could be produced for specific system constraints for real-world problems. For example, missile defense or target classification requires real-time classification. Space applications also have restrictions such as computation power and memory, due to the radiation-hardening constraint on spaceflight processors and systems. Such problems cannot always leverage traditional ensemble learning and conventional machine learning algorithms.

Team learning introduces social interaction, team-building, and knowledge transfer that are inherent in human teams. In order to achieve true team-based learning and collaboration, models are needed for how humans cognitively learn, interact, and teach one another. New and novel machine learning methods are also needed for information sharing and collaboration. Incorporating concepts such as trust, confidence, and negotiation, along with using them to intelligently reason to arrive at a collective decision, may allow for higher accuracy classification or task performance. Investigating other next-generation learning architectures, such as Deep Learning (incorporate both substance and meaning) and hierarchical systems, is also important to advance the state-of-the-art in multi-agent machine learning.

Acknowledgments

Author C.M. Gifford performed this research while part of the Center for Remote Sensing of Ice Sheets (CReSIS) at the University of Kansas. The authors would like to thank Anthony Hoch at the CReSIS for helpful discussions about radar theory, and assistance in constructing the data set for this study. This material is based upon work supported by the National Science Foundation under Grant No. ANT-0424589. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


