A PRELIMINARY STATISTICAL ASSESSMENT OF INSTABILITY PREDICTABILITY IN OBSERVED SNOW PROFILES

Sondre Wold

Norwegian Water Resources and Energy Directorate (NVE), Oslo, Norway

ABSTRACT: Observational recordings of snow conditions help avalanche forecasters assess snowpack instability over large areas. In Norway, trained observers submit snow profiles and Extended Column Test (ECT) results through the RegObs platform as part of the national avalanche forecasting service. However, observations collected under field conditions contain unavoidable noise, raising the question of how much information about snow instability is preserved in recorded snow profiles. Using six seasons of observational data from Norway, we investigate this question using ECT outcome as a proxy for snow instability. We train a suite of machine learning classifiers to predict ECT outcome from recorded profiles, using classifier performance as an empirical estimate of the predictive signal in the data. We find that models that process the snowpack layer-by-layer outperform models based on profile-level aggregates, showing that the layered structure of the snowpack carries information that can be lost through aggregation. We also provide an estimation of the association between different grain types and ECT outcome using the pointwise mutual information.

KEYWORDS: instability tests, sequence modeling, data validation, machine learning

1. INTRODUCTION

Assessing snow instability is an important part of avalanche forecasting. To accurately make this assessment, forecasters typically draw on multiple sources of information, such as meteorological data, observational reports and historical data on the development of the snowpack. Automated prediction of snow instability based on one or more of such sources can aid the forecasting process, and previous research has shown that model-based predictions of both avalanche danger and snow instability now achieve sufficient accuracy to offer practical value to forecasters (Techel, Helfenstein, et al. 2024). A notable example of an operational tool is the p_{unstable} estimation introduced in Mayer et al. (2022), which is based on a random forest classifier trained on simulated snow stratigraphy produced by the physics-based SNOWPACK model (Bartelt and Lehning 2002).

While data-driven methods, such as those based on machine learning, are naturally apt for developing predictive tools, they can also be used to obtain an empirical estimate of the information content inherent in a given dataset. In this paper, we predominately take this approach. Using both traditional statistical methods and techniques based on deep learning, we conduct a preliminary study of statistical predictors of snow instability in observational data. Based on six seasons of observational data from Norway, we address two research questions: First, how much information about snow instability is gained by knowing the grain types present within the snowpack? Second, are models that process the profiles layer-by-layer better for predicting ECT outcome compared to models that process the whole profile as summary features?

For both questions we use the outcome of the extended column test (ECT) (Simenhois and Birkeland 2006) as a proxy for snow instability. Our analyses draw on 5 932 observations from the 2020-2026 winter seasons in Norway, with each sample including both an ECT test result and a recorded snow profile, all conducted by observers with documented training. In this preliminary work, our primary goal is to validate observational data using a data-driven approach, which we hope can be extended in the future to include the development of robust prediction models for snow instability.

We begin by describing the data collection procedure and methodological framework. We then analyze the information content of grain types, addressing our first research question. To address the second, we present a series of machine learning experiments and evaluate the ability of different models to predict ECT outcome from observed profiles. In particular, we investigate whether models specifically developed for sequential data outperform models based on profile-level features. Through ablation experiments, we also gauge the contribution of single features to the overall predictive performance. We end with a discussion on our findings and their limitations.

* *Corresponding author address:*
Sondre Wold, NVE
Middelthuns gate 29, 0368 Oslo, Norway

https://doi.org/10.15788/1790098791

2. DATA COLLECTION

We use the openly available RegObs API provided by NVE, collecting all observations from the 2020-2026 winter seasons, with each season running from the beginning of December to the end of May. In order to retain only high-quality observations, we conduct several filtering steps. As RegObs accepts observations from the general public, we exclude any observations by accounts with no documented training in recording snow profiles.[1]

To exclude homogeneous snowpacks, which are common at the beginning and end of the season, we only include profiles with three or more layers (median=7). Homogeneous packs are often stable, either due to insufficient slab formation early in the winter or a largely settled snowpack in late spring. We further restrict the dataset to observations containing an ECT result recorded at the same GPS location as the snow pit. When multiple ECT results are available for a single pit, we select one at random.

As a proxy for instability, ECT outcomes are grouped into a binary target variable, where ECTP and ECTPV are classified as unstable and ECTN and ECTX as stable. Although the raw data also contain the number of loading steps required to initiate failure, we do not use this information. Previous studies have shown that the primary information gained from an ECT is whether fracture propagation occurs rather than the loading score itself (Simenhois and Birkeland 2009; Techel, Winkler, et al. 2020).[2] We therefore define instability solely in terms of propagation outcome. This approach is also supported by the originally proposed interpretation of ECT results, which treated propagation as an indicator of instability and non-propagation as an indicator of stability, as discussed in Toft et al. (2024).[3]

An overview of the collected data and the target variable distribution is shown in Table 1. After filtering, we have 5 932 total samples, each containing a recorded snow profile and an ECT result. The target variable is significantly imbalanced, with only 38.3% of tests containing an unstable result.

[1]This corresponds to only keeping observations from observers with a three star rating or more in RegObs, indicating completion of formal training in snow profile observation.

[2]See Birkeland et al. (2023) for a broader comparison and discussion of instability tests.

[3]According to Toft et al. (2024), an ECTX result (no fracture after all the loading steps are completed) was to be regarded as indecisive. In our work we decided to regard this as part of the stable category.

Season	N	Unstable (%)
2020-2021	661	45.7
2021-2022	769	38.6
2022-2023	974	34.6
2023-2024	1101	38.7
2024-2025	1167	34.4
2025-2026	1260	40.6
Total	5932	38.3

Table 1: Dataset summary after filtering.

3. INFORMATION CONTENT OF GRAIN TYPES

Before developing any predictive models (see Section 4.), we begin by quantifying the marginal association between individual grain types and ECT outcome. The grain type classification of each layer is indicative of the different processes occurring in the snowpack. Furthermore, it is naturally associated with other variables recorded, such as hand hardness and grain size. Hence, the grain type classification is a natural choice for validating that field recordings carry information about instability despite the inherent noise associated with the recording procedure, such as adverse weather conditions. By analyzing the marginal association between grain types and ECT outcome, we essentially probe whether or not there is at least *some* statistical rationale for developing a predictive model on the same data.[4]

Our analysis gives a simple estimate of how much *information* about instability is contained in the presence of specific grain types, independent of their position or thickness in the profile. To quantify these associations, we calculate the Pointwise Mutual Information (PMI) (Church and Hanks 1990), which measures the strength of association between two events. PMI expresses association strength in bits, is symmetric with respect to the direction of association, and equals zero when two events are independent, making it straightforward to interpret. In our case, the PMI between the presence of a specific grain type in a snow profile, for example depth hoar, and ECT outcome, measures how much more likely instability is given that the profile contains depth hoar, relative to the two events being independent.

We bootstrap the PMI calculation by resampling our dataset 5000 times with replacement, producing a confidence interval that reflects the uncer-

[4]And conversely, if no marginal association is found, this would be highly indicative of some source of error in the process of recording the grain types that would exist across observers.

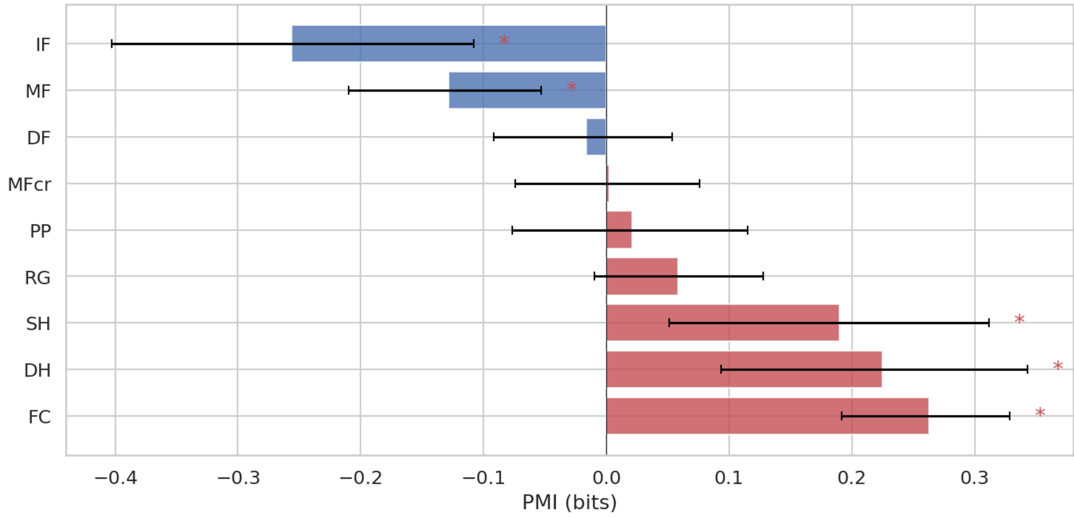

Figure 1: The PMI between the profile-level presence of a grain type and the ECT outcome as binary proxy for snow instability. The x-axis denotes the PMI as bits, while the y-axis denotes the grain type. The * symbol indicates a significant result. Only grain types with prevalence $\geq 5\%$ shown. Confidence intervals are obtained from 5000 bootstrap resamples. A PMI of 1 bit indicates that the grain type and ECT outcome co-occur twice as frequently as expected under statistical independence.

tainty in our measurements given the dataset. We consider the association significant when the 95% confidence interval (CI) excludes zero, indicating that the calculated association is unlikely to be due to chance.[5]

Figure 1 shows the PMI between the presence of different grain types in a profile and ECT outcome. A positive PMI, as measured in bits, indicates an association between the grain type and getting an unstable ECT result, while a negative PMI indicates a negative association, meaning that the grain type is associated with a stable ECT result. Our results show that the presence of grain types typically regarded as indicative of an unstable snowpack indeed have a higher PMI in our sample. For example, facets, depth hoars and surface hoars all have a significant positive association with an unstable ECT result. Likewise, grain types typically associated with a more stable pack display a significant negative association, such as melt forms and ice formations. The presence of rounded grains, precipitation particles, melt-freeze crusts and decomposing particles show no significant association in either direction, meaning that establishing the presence of either grain type does not enable us to confidently distinguish the result from an independent relationship between the two events.

[5]This means that the 2.5th (CI lower) and 97.5th percentile (CI upper) both are on the same side of zero.

4. PREDICTING ECT OUTCOME

The PMI analysis measures marginal associations between individual grain types and ECT outcome. However, a strong marginal association does not guarantee that a variable adds predictive value once other correlated features are available. If hand hardness already captures much of the same information, for example, grain type may contribute little beyond what is already known. As a next step, we therefore evaluate whether combinations of features can be used to predict instability at the profile level. This experiment also allows us to compare models that use fixed profile-level summaries with models that operate directly on the ordered layer sequence, which we hypothesize are better suited to capture the nature of snow stratigraphy, which has a natural ordering of layers from the surface to the ground.

4.1. Models and input features

We compare six machine learning classifiers with increasing model complexity. The first group of models uses seven profile-level summary features: mean hand hardness (transforming the traditional classification of hardness to an ordinal scale), hardness range (min and max in the profile), number of layers, weak layer fraction, maximum hardness jump between consecutive layers, number of critical interfaces, and a binary indicator for the presence of a hard layer immediately over a soft layer. The weak layer fraction is defined as the proportion of layers containing facets, depth hoar or surface hoar. A critical interface is defined as a hardness increase of at least two steps between adjacent layers, for example going from F to 1F.

We did not include the profile temperature gradient or the grain size of each layer due to a large degree of missing data. By including these variables, especially grain size, we would effectively reduce the dataset beyond what is typically needed in order to run machine learning models of a certain complexity and parameter size.

The majority-class baseline predicts the most common label in the training data and serves as a lower-bound reference. As instances of simpler machine learning models, we use both a logistic regression model and a random forest classifier. We also train a simple feed-forward neural network (MLP) on the summary features, providing a neural baseline.

The second group of models operate directly on the ordered snow profile. Each layer is represented by hand hardness, layer thickness, depth from the surface, and grain type. Hand hardness is encoded as an ordinal variable, while grain type is represented by an embedding learned during training. The model receives the full profile as a sequence of layers ordered from the surface downward. We evaluate two sequence models: a 1-dimensional convolutional neural network (CNN) (LeCun et al. 1989) and a Transformer encoder (Vaswani et al. 2017) in the style of Devlin et al. (2019). Both models produce representations of each layers in sequence, and these are then reduced via mean pooling and passed to a linear classification head. Both models are trained with padding to handle the different number of layers in each profile. For the Transformer, the sequential ordering is encoded using a learned positional embedding, while for the CNN, this is enforced as the convolutions are applied to local, ordered (surface to ground) neighborhoods.

The features selected for both model groups are motivated by the results of the preceding PMI analysis and general intuition. As such, they do no represent an ideal feature set, and the results should be interpreted accordingly. As a preliminary work, it is beyond the scope of this study to explore a large set of features with the aim of maximizing predictive performance; as stated earlier, our primary goal here is to probe the predictive signal of some variables in the snow profiles w.r.t. instability, with the aim of developing more robust predictive models in the future.

4.2. Training and evaluation

We evaluate all models using leave-one-season-out cross-validation. In each fold, one full winter season of data is held out as the test set, while the remaining five seasons are used for training. The model is trained from scratch in each fold and evaluated on the held-out season. This procedure is repeated once for each season, and the resulting scores are averaged. We chose this scheme to reflect the intended use case, since, in forecasting, a model will generally be applied to future seasons that were not available during training. For the MLP, CNN, and Transformer, we select training hyperparameters using a nested procedure based on the approach outlined in Varma and Simon (2006).

4.3. Evaluation metric

We report the performance of our models using the area under the receiver operating characteristic curve (AUC-ROC). Each classifier returns a continuous score for each profile, interpreted as the model's confidence that the profile is unstable. AUC measures how well these scores rank unstable profiles above stable profiles. A perfect ranking has an AUC of 1, while a random ranking has an AUC of 0.5, despite the fact that the dataset is imbalanced with only 38.3% of the profiles being unstable.

5. Results

Our results can be seen in Figure 2. Generally, we observe that the CNN and Transformer models outperform the models based on profile-level summary features. The CNN obtains the highest mean AUC, with a score of 0.719, followed by the Transformer with an AUC of 0.705. In comparison, the aggregate classifiers all score around 0.65 AUC. The difference between the best sequence model and the best aggregate model is 0.07 AUC points, indicating that the ordered layer sequence contains predictive information that is lost when profiles are reduced to summary statistics.

In Table 2, we show the results of ablating different subsets of features for the Transformer model. Interestingly, although perhaps not too surprisingly, the contribution of grain type depends on the other available input variables in our experiments. While excluding the grain type classification from the feature set causes a drop in performance of 0.006 AUC, the exclusion of hand hardness causes a 0.03 AUC drop. This indicates that the grain type classification in our data carries predictive information about instability, but that much of this in-

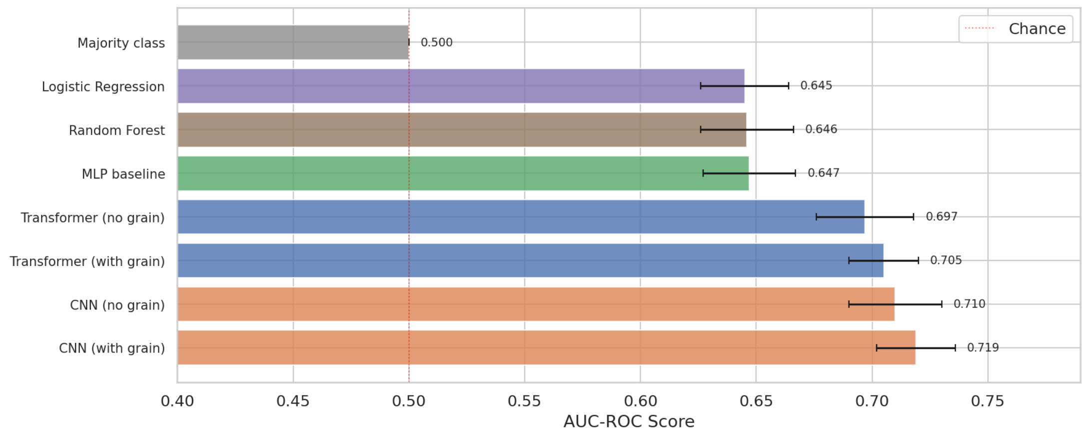

Figure 2: The AUC-ROC score of the different tested models on the task of predicting binary ECT outcome given snow profile information. The plotted score is the mean score over all tested seasons, together with the standard deviation. For the sequence models we also compared including or excluding grain type classification as a feature.

Feature Set	AUC
All features	0.708 $\pm$0.016
Leave out: grain type	0.702 $\pm$0.016
Leave out: layer thickness	0.701 $\pm$0.010
Leave out: depth to surface	0.700 $\pm$0.015
Hardness + grain type	0.697 $\pm$0.017
Leave out: hardness	0.678 $\pm$0.027
Hardness only	0.676 $\pm$0.020
Grain type only	0.663 $\pm$0.024

Table 2: Feature ablation results for the Transformer model evaluated using leave-one-season-out cross-validation. Values are mean AUC $\pm$ standard deviation across folds. Same settings as in Figure 2, but a different run.

formation overlaps with what is already captured by other features, especially hand hardness.

6. DISCUSSION

Our results show that observed snow profiles contain a measurable, but limited, signal about ECT outcome. For example, grain type is clearly associated with instability in our PMI analysis: facets, depth hoar and surface hoar are positively associated with unstable tests, while melt forms and ice formations are negatively associated. This supports the view that the observational data contain physically meaningful information, and that snow profiles reflect expected physical relationships despite the challenging conditions under which they are typically collected. We believe the contribution of this analysis is best understood as a data-driven method for validating data quality.

In our second experiment, we find that grain type adds no significant additional predictive value for our sequence models when other layer variables are included. This does not imply that grain type is unimportant for snow stability assessment, but suggests that, for predicting binary ECT outcome in this dataset, much of its signal overlaps with variables that are simpler to record consistently in the field, such as hand hardness. As hand hardness and grain type are highly dependent variables, this is to be expected, but it shows that for the task of predicting ECT outcome, hand hardness suffice.

Numerically, our strongest result is that sequence models outperform models based on profile-level summaries. Preserving the ordered layer structure improves prediction, while increasing classifier complexity on aggregate features does not. Even so, the best model performance, achieved by the CNN, remains below what would be required for stand-alone operational use. Our models are best viewed as tools for estimating information content in observational profiles, or as weak auxiliary signals in a broader forecasting workflow. In a similar study, Pérez-Guillén et al. (2025) developed a machine learning model with $\sim$ 70% agreement with forecasters on the avalanche danger level, and they come to a similar conclusion, that such models are best viewed as a good second opinion. Nevertheless, we believe we have shown that a machine learning model that caters to the sequential nature of snowpack stratigraphy is better suited than profile-level models, a finding that we believe can be used in future work to advance already existing approaches to the operationalization of predictive models.

7. LIMITATIONS

The primary limitation of this study is the use of ECT outcome as a proxy for snow instability. While ECT results are widely used as indicators of instability, they represent the response of a single test at a specific location. The dataset is also subject to uncertainty inherent to field observations. Both snow profile recordings and ECT outcomes depend on observer judgment and expertise and are collected under varying field conditions. Such variability is expected to introduce noise into both the predictor and target variables. Some of the observers are also highly frequent reporters to the NVE RegObs database, and previous work has shown that there are individual differences in parts of the execution of the ECT (Toft et al. 2024), implying that there could be observer-specific confounding effects.

By design, our models use only information from the recorded snow profile and do not incorporate terrain variables (aspect, elevation, slope angle) or meteorological context. This choice isolates the predictive signal in profile observations alone, but implies that our reported performance represents a lower bound on what could potentially be achieved by incorporating other features in the prediction pipeline.

We also want to emphasize that this is preliminary work. Although our work indicates an increase in performance in the sequence models, and we try to control for some confounders, there is always the possibility of other model-specific confounding effects. For example, the sequence models differ from the simpler models in multiple ways, not only the sequential encoding. To attribute the increase in performance to the sequential mechanism alone is therefore difficult. As always, such comparisons should therefore be interpreted conservatively.

Finally, our evaluation only assesses generalization across seasons within the Norwegian observation network. Further work is needed to determine whether the observed relationships are transferable to other snow climates, regions, or observer networks.

REFERENCES

Bartelt, P. and Lehning, M. (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part I: numerical model. In: Cold Regions Science and Technology 35.3, pp. 123–145. ISSN: 0165-232X. https://doi.org/https://doi.org/10.1016/S0165-232X(02)00074-5.

Birkeland, K. W., Herwijnen, A. van, Techel, F., Bair, E. H., Reuter, B., Simenhois, R., Jamieson, B., Marienthal, A., Chabot, D., and Schweizer, J. (2023). “Comparing stability tests and understanding their limitations”. In: Proceedings of the 2023 International Snow Science Workshop, Bend, OR, http://arc. lib. montana. edu/snow-science/item/2855 (last access: 15 August 2024).

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. In: Computational Linguistics 16.1, pp. 22–29.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Ed. by Burstein, J., Doran, C., and Solorio, T. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems 2.

Mayer, S., Herwijnen, A. van, Techel, F., and Schweizer, J. (2022). A random forest model to assess snow instability from simulated snow stratigraphy. In: The Cryosphere 16.11, pp. 4593–4615. https://doi.org/10.5194/tc-16-4593-2022.

Pérez-Guillén, C., Techel, F., Volpi, M., and Herwijnen, A. van (2025). Assessing the performance and explainability of an avalanche danger forecast model. In: Natural Hazards and Earth System Sciences 25.4, pp. 1331–1351. https://doi.org/10.5194/nhess-25-1331-2025.

Simenhois, R. and Birkeland, K. W. (2006). “The extended column test: a field test for fracture initiation and propagation”. In: Proceedings ISSW. Vol. 10, pp. 79–85.

Simenhois, R. and Birkeland, K. W. (2009). The extended column test: test effectiveness, spatial variability, and comparison with the propagation saw test. In: Cold Regions Science and Technology 59.2-3, pp. 210–216.

Techel, F., Winkler, K., Walcher, M., Herwijnen, A. van, and Schweizer, J. (2020). On snow stability interpretation of extended column test results. In: Natural Hazards and Earth System Sciences 20.7, pp. 1941–1953. https://doi.org/10.5194/nhess-20-1941-2020.

Techel, F., Helfenstein, A., Mayer, S., Pérez-Guillén, C., Purves, R., Ruesch, M., Schmudlach, G., Soland, K., and Winkler, K. (2024). “Human vs Machine–Comparing model predictions and human forecasts of avalanche danger and snow instability in the Swiss Alps”. In: Proceedings International Snow Science Workshop, Tromsø, Norway, pp. 23–29.

Toft, H. B., Verplanck, S. V., and Landrø, M. (2024). How hard do avalanche practitioners tap during snow stability tests? In: Natural Hazards and Earth System Sciences 24.8, pp. 2757–2772.

Varma, S. and Simon, R. (2006). Bias in Error Estimation When Using Cross-Validation for Model Selection. In: BMC Bioinformatics 7, p. 91. https://doi.org/10.1186/1471-2105-7-91.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In: Advances in neural information processing systems 30.

A REPRODUCIBILITY STATEMENT

Data and experimental code are made publicly available .[6] We provide both zip archives of the

[6] `https://github.com/SondreWold/snow-stability-prediction-issw`

data as well as the possibility to re-download the raw data from RegObs. Changes in the API might break consistency in the raw data, reproduction must therefore be conducted on the provided zip archives. All experiments ran on a single NVIDIA RTX PRO 500 GPU.

B IMPLEMENTATION

The simpler, aggregate models were used out-of-the-box from the `sklearn` library, while the neural networks were implemented using `PyTorch`.