**Introduction**

Avalanche forecasting integrates mountain meteorology, snowpack physics, terrain analysis, and risk communication to reduce loss of life and support safer winter travel. In Scotland, the Scottish Avalanche Information Service (SAIS) publishes daily next-day hazard assessments across six high-mountain regions during winter which draws upon field observations, snowpack tests and weather guidance (Scottish Avalanche Information Service [SAIS], n.d.). This report contributes a data-driven complement to that long-running operational effort. Our task is defined as to construct and evaluate a neural network model that predicts the forecasted avalanche hazard (FAH).

This research isn’t meant to replace expert judgment. It gives forecasters a second opinion that consistently flags risky combinations of weather, terrain, and snowpack variables. In Scotland’s maritime, wind-dominated winters with frequent storm cycles and rapid thaw freeze shifts this kind of decision support sharpens situational awareness and strengthens risk communication (SAIS, n.d.; EAWS, n.d.).

**Background**

Scotland’s avalanche forecasts are produced by the Scottish Avalanche Information Service (SAIS). The SAIS conducts daily fieldwork, snow profiles, stability tests and terrain analysis that is blended with meteorological guidance to issue next-day public danger ratings, snowpack summaries and travel advice. This information spans Torridon, Northern Cairngorms, Southern Cairngorms, Creag Meagaidh, Lochaber, and Glencoe. The communication follows the standard EAWS danger scale to keep messages consistent and comparable.

Avalanches are rapid snow flows that can contain ice, rock, and vegetation. The most dangerous to travellers are slab avalanches, where a cohesive slab slides on a weaker layer once shear strength is exceeded by gravitational and external loading. The triggers for an avalanche may be natural (snowfall, wind loading, warming) or manmade (a skier, climber, or snowmobile). These avalanches most often occur on slopes around 34–45 degrees. In Scotland’s maritime, wind-dominated winters wind slabs are common (Schweizer, Jamieson, & Schneebeli, 2003).

Avalanche forecasting looks to improving consistency, calibration, and timeliness of hazard assessments which can essentially save lives and reduce societal costs. In Europe, despite substantial growth in backcountry participation, the long-term average annual avalanche fatality count has remained broadly steady at 100 per year across the Alps. This is attributed in part to better education, equipment, and forecasting (Techel et al., 2016). Scotland’s totals are smaller, but fatal and non-fatal involvements recur most winters and SAIS seasonal reports routinely document hundreds of observed avalanches (SAIS, n.d.). The impacts however extend beyond casualty numbers as they tend to disrupt transport, tourism, and emergency services.

**Dataset**

The dataset is an operational archive produced by the Scottish Avalanche Information Service (SAIS) over approximately fifteen winter seasons reflecting real-time fieldwork. The records are created by forecasters working in severe-weather environments and under time constraints hence the data naturally exhibits uneven sampling across storm cycles and regions with also missing entries during access-limited periods. The variables available in the dataset contain physically informed features inherent within the environment as defined below:

* Terrain susceptibility (*Incline, Aspect, Alt*)

Slope angle is the primary release control for slab avalanches as most events initiate in the mid-30s to mid-40s degrees so *Incline* is a direct indicator of where failures are most likely. *Aspect* governs how much wind and sun a slope gets, when analysed with recent *Wind.Dir* it tells you which slopes are lee and windward. *Alt* defines the altitude of the observation site in meters above sea level. (McClung & Schaerer, 2006; Schweizer, Jamieson, & Schneebeli, 2003).

* Loading and weather drivers (*Air.Temp, Wind.Dir, Wind.Speed, Cloud, Precip.Code, Drift, Rain.at.900, Summit*)

Near-surface *Air.Temp* shows warm spells that weaken bonds, while *Wind.Dir* and *Wind.Speed* together determine snow transport and where slabs quickly build on lee and cross-loaded slopes. *Cloud* cover explains cooling/heating that drive near-surface faceting or slows refreezing. *Precip.Code* indicates the type and intensity of new loading (snow, showers, etc). *Rain.at.900* flags the rainfall at 900m altitude affecting the snowmelt. Field *Drift* observations confirm if the snow has been blowing recently. *Summit****.\**** variables indicate to us what is happening at the Summit where most loading starts.(EAWS, n.d.; McClung & Schaerer, 2006; Schweizer et al., 2003).

* Snowpack structure and stability proxies (*Total.Snow.Depth, Foot.Pen, Ski.Pen, Max.Temp.Grad, Max.Hardness.Grad, Snow.Temp, Wetness, Crystals, Snow.Index, No.Settle, AV.Cat, Insolation*)

*Total.Snow.Depth* reflects total depth of snow cover at the observation site, while *Foot.Pen/Ski.Pen* indicate near-surface strength. *Max.Temp.Grad* signals the temparture across the snowpack indicating instability. *Max.Hardness.Grad* captures hardness within the snowpack reflecting the layer differences. *Snow.Temp* defines the temperature within the snowpack and *Wetness* defines the degree of snow wetness. *Crystals* identify type or size of snow crystals observed which influences avalanche risk. *Snow.Index* provide cues on the stability of the snow. *AV.Cat* classifies avalanches by type/severity*. Insolation* measures the impact of solar radiation affecting snow melt*.* (EAWS, n.d.; Schweizer et al., 2003; McClung & Schaerer, 2006).

By relating these predictors to FAH the model can learn recurrent, non-linear patterns. It highlights situations where higher hazard is more likely, so forecasters can focus their checks and make cleaner day-to-day calls (EAWS, n.d.; SAIS, n.d.).

**Methodology**

This research builds a supervised, ordinal-aware classification pipeline to predict next-day Forecast Avalanche Hazard (FAH) from the operational SAIS records. The end-to-end workflow was implemented in R and organized into clearly defined stages—data auditing, preprocessing, feature engineering, imputation, encoding, feature selection, data splitting and class rebalancing, neural-network modeling, and evaluation. Reproducibility is enforced via *set.seed(42)* and some of the core libraries include *tidyverse* for data wrangling, *caret* for preprocessing/splitting, *glmnet/randomForest* for feature selection signals, *ROSE* for class balancing, and *keras3/tensorflow* for model training.

We used large language models (LLMs) to accelerate routine coding tasks. Each team member supplied an LLM with a structured brief (data schema, variable descriptions, target definition, leakage constraints, evaluation metrics) to draft boilerplate R code for preprocessing, imputation functions, model scaffolding (keras3), and plotting. To promote diversity of ideas and reduce single-model bias, different teammates intentionally used different LLMs. This use of LLMs sped up drafting, but final code, methodological choices, and validation remained human-curated.

**EDA**

**Subsection: Target Variable Distribution**

A graph of different colored squares

AI-generated content may be incorrect. A pie chart with text on it

AI-generated content may be incorrect.

From Figures X and Y above, we observe that the FAH target variable exhibits severe class imbalance, with safer classes dominating the distribution: Low (≈32.5%) and Moderate (≈30.7%) together account for ≈63% of observations, while Considerable- is moderately represented (≈23.6%). In contrast, the higher-risk classes—Considerable+ (≈8.8%) and especially High (≈4.4%)—are rare. This skew implies that a model could achieve high overall accuracy by favoring common classes while failing to detect infrequent but critical high-hazard events. To mitigate this, we employ ROSE balancing during training and prioritize macro-averaged metrics alongside high-risk class recall in evaluation.

**Subsection: Missing Data Analysis**

A graph of a number of objects

AI-generated content may be incorrect.

From Figure X above, the missing data diagnostics reveal that AV\_Cat and Ski\_Pen exhibit the highest gaps (≈20%+), prompting their removal from the dataset during preprocessing. Following these, a cluster of snowpack microstructure and condition variables (e.g., Crystals, Wetness, No\_Settle) show moderate missingness, alongside certain wind-direction and temperature fields. The target variable *FAH* has negligible missingness, likely due to operational teams making inferences when weather conditions do not allow for some metrics to be captured.

**Subsection: Temporal Analysis**

A group of blue and red graphs

AI-generated content may be incorrect.

From the figure x above that denotes temporal coverage we can see that the archive spans many winters with stable annual volumes after the early years and a strong seasonal concentrationin Dec–Mar, tapering in shoulder months and near-zero in summer (as expected). The daily series shows a burst of entries followed by lulls which tell us that data arrives in clusters during winter storm cycles and thin out outside winter. These patterns justify including month/season features and caution against assuming stationarity across years.

A graph of a bar

AI-generated content may be incorrect.

From Figure X above, which analyses FAH versus OAH pairs as a verification benchmark for forecast reliability, we observe an overall accuracy of approximately 70–75%. Per-class performance varies notably: accuracy is highest for Low risk (mid-90s%), followed by Moderate (70%), then decreases for Considerable- (60–70%), drops substantially for Considerable+ (35%), and improves slightly for High (~45%). This pattern suggests that forecasters—and by extension, predictive models—struggle most with distinguishing adjacent hazard boundaries. Accordingly, our model evaluation emphasizes ordinal metrics such as adjacent accuracy (±1 level), critical-miss rate (underpredicting high risk), and high-risk sensitivity to ensure practical utility in avalanche forecasting.

The EDA completed paints a clear operational picture with our target strongly skewed toward Low/Moderate days, so hence a model judged on plain accuracy could look good while failing on the rare but consequential upper hazards. In the implementation, we therefore look to rebalance the training split with ROSE and commit to macro metrics and high-risk recall at evaluation. The missingness is structured and not entirely random, primarily with the >20% gaps in *AV.Cat* and *Ski.Pen* which we drop. The clusters in snowpack detail fields and some wind directions reinforce the choice of *Area×Season* imputation (fit on train only) and the inclusion of seasonal features. The temporal profile confirms winter-centric usage with stable annual volumes, motivating Month/Season encodings and caution about year-to-year drift. Finally, FAH–OAH verification (~70–75% overall, strongest on Low, weakest in the Considerable+/High tail) mirrors the hardest real-world boundaries, hence we will look to prioritize adjacent accuracy (±1), critical-miss rate, and high-risk sensitivity to ensure the model not only performs well on average but reliably catches dangerous days**.**

**Data Processing**

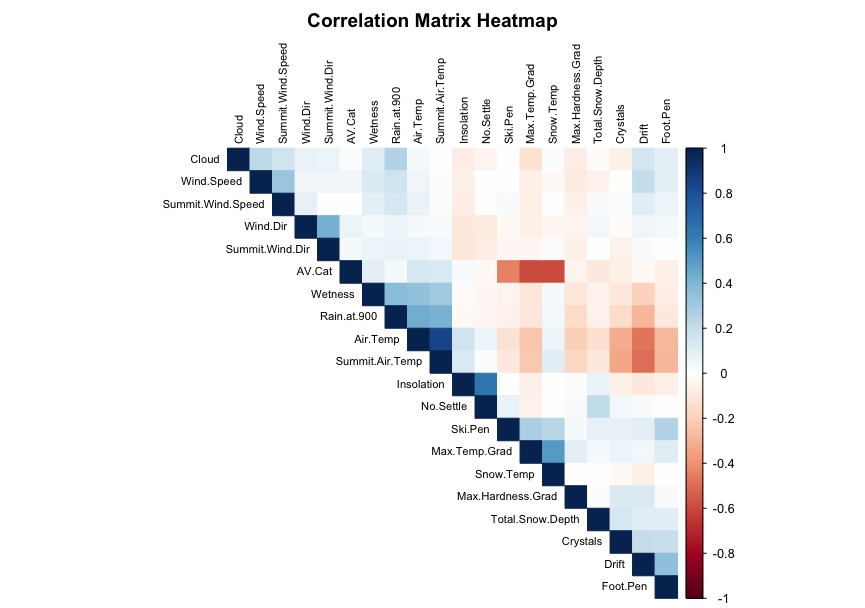
**Missing Data I**

A diagram of different types of data

AI-generated content may be incorrect.

From the figure x above the boxplots show two kinds of outliers. Some are implausiblevalues with *AV.Cat* having huge positive/negative magnitudes and >20% missingness. *Cloud* occasionally exceeds 100% or goes negative. The others may be described as legitimate extremes as *Snow.Index* has a heavy right tail, Crystals has rare large values amid many zeros, and *Rain.at.900* behaves like a binary indicator (0/1), so the “1”s only look like outliers on a numeric boxplot. Latitude spans the expected Scottish band and is not a true outlier.

Guided by this, we used a two-tier outlier strategy by first applying a physical-plausibility filters to drop *AV.Cat*, cap *Cloud* to [0,100], discard *Aspect* > 360°, *Incline* <0 or >90°, and remove negative or absurd readings for wind, snow depth, and penetrations. Second, for variables where extremes are meaningful we used robust IQR trimmingwith a wide cutoff (k≈3–3.5) to prune only extreme spikes while preserving genuine storm extremes. We also recast *Rain.at.900* as a binary feature to avoid mislabelling valid “1”s as outliers. This approach reduces measurement/error noise, stabilizes scaling and NN training.



The figure x above notes the correlation heatmap that indicates coherent but mostly moderate blocks. It is noted that the temperatures (*Air.Temp, Summit.Air.Temp*) co-vary while *Cloud* and *Insolation* are opposing as expected. Wind speeds correlate across levels with noisy directional correspondence and penetration/strength (*Foot.Pen, Ski.Pen*) associate with *Total.Snow.Depth* *Max.Temp.Grad* and *Max.Hardness.Grad* also relate to temperature and snow-structure variables in a non-simple way. Altogether, this signals multicollinearity with strong overlap among thermodynamic and wind features, this can upset linear models and slow neural-network training. To reduce redundancy, we used ensemble feature selection (LASSO, random forest importance, and correlation filters). We also encoded Aspect with sine and cosine so the model handles direction smoothly without a jump at 0°/360°.

Building on those patterns, we engineered features that map avalanche mechanics into stable, low-redundancy inputs:

* **Wind–cold interaction**: a simple *wind-chill* proxy (*Air.Temp − 0.6×Wind.Speed*) to reflect near-surface cooling.
* **Vertical stability**: *summit–valley temperature gradient* (*Summit.Air.Temp − Air.Temp*) asa rough indicator of atmospheric stratification/transport potential
* **Snow–altitude interaction**: *Total.Snow.Depth × (Alt/1000)* to allow depth effects to vary with elevation (rain/snow line).
* **Circular aspect encoding**: cos(Aspect) and sin(Aspect) to preserve directionality without discontinuity at 360°/0°.
* **Seasonality**: calendar *Month*, *Day\_of\_Year*, and *Season* factors to capture intra-winter cycles.

## **Implementation**

1. **Data Splitting and ROSE Balancing**

The cleansed and adapted dataset of 7530 observations was split into training (70%), validation (15%), and test (15%) subsets, stratified by hazard level to preserve class proportions. The initial split produced 5,273 training samples, 1,129 validation samples, and 1,128 test samples, corresponding to the 7,530 complete cases used in the modelling process. The training set mirrored the imbalance present in the data, with Low (class 0) and Moderate (class 1) hazards each exceeding 1,700 samples, while High (class 4) had only 145 cases. Both validation and test sets reflected similar proportions, preserving the scarcity of higher hazard levels.

This distribution represented an imbalance ratio of approximately 12.4:1 between the largest (Low) and smallest (High) classes in the training data, a level of imbalance that would severely bias any classifier toward predicting majority classes.

To address this, the ROSE was applied exclusively to the training data after splitting, leaving validation and test sets untouched to preserve their natural distributions for unbiased evaluation. Because ROSE natively handles binary classification, an iterative one-versus-rest strategy was implemented. Each minority class was balanced against the remaining classes, generating synthetic examples until minority classes approached 95% of the majority class size. This approach avoided artificially perfect balance while still substantially improving minority class representation.

The training set expanded from 5,273 to 8,641 samples, with each hazard level represented by roughly 1,700 cases. The balancing reduced the imbalance ratio to 1.05:1, corresponding to a 91.5% reduction in class imbalance. This substantial improvement provided the neural network with sufficient minority class representation while preserving realistic proportions in validation and test datasets for fair evaluation.

1. **Neural Network Architecture**

After hyperparameter tuning across three candidate configurations, the optimal network achieved a validation accuracy of 0.6023. The final architecture consisted of three hidden layers (384, 192, 96 units) with progressively increasing dropout (0.25, 0.35, 0.45) and Gaussian noise regularization to control overfitting. The network used the Adam optimizer with a learning rate of 8e-04, a batch size of 48, and trained for 53 epochs.

| **Layer** | **Units** | **Dropout** | **Notes** |
| --- | --- | --- | --- |
| Input | – | – | 25 selected features |
| Hidden Layer 1 | 384 | 0.25 | Dense + BatchNorm + GaussianNoise |
| Hidden Layer 2 | 192 | 0.35 | Dense + BatchNorm + Dropout |
| Hidden Layer 3 | 96 | 0.45 | Dense + BatchNorm + Dropout |
| Output | 5 | – | Softmax (multiclass, ordinal) |

*Table 4: Final neural network architecture*

The training and validation curves showed that the model achieved stable convergence, though validation accuracy plateaued around 0.60, reflecting the intrinsic difficulty of predicting the rare high-risk classes.

**Model Performance Evaluation**

**Overall Performance**

The final evaluation of the neural network was conducted on the held-out test dataset (n = 1,128), which was not used during training or validation. This ensured that the reported results represent an unbiased estimate of the model’s real-world generalization ability. The network achieved an overall test accuracy of 61.2% with a corresponding test loss of 0.978. This represents a substantial improvement over the baseline accuracy of 33.7%, which corresponds to a naive strategy of always predicting the most frequent class (Low). In relative terms, the neural network provided an 81.6% improvement over this baseline, demonstrating that it captured meaningful patterns in the meteorological, geographical, and snowpack predictors beyond class frequency distributions.

The EDA findings help explain both the strengths and limitations of this performance. During data exploration, the FAH distribution was found to be highly imbalanced, with Low and Moderate hazard levels dominating the dataset and the High category being very rare. Even after balancing the training set through ROSE sampling, the validation and test sets retained this natural skew. This explains why the model excelled at predicting common categories but struggled with rarer ones such as Considerable+ and High, as reflected in class-level metrics. Furthermore, the geographical analysis showed that certain areas like Lochaber and Torridon had consistently more observations than other regions, while seasonal analysis revealed strong clustering of High hazard levels during winter peaks. These patterns suggest that the network learned to associate common seasonal and spatial features with Low/Moderate levels more reliably than with rarer, extreme events, which limited absolute performance.

To provide a more nuanced view of model performance across imbalanced classes, multiple complementary metrics were calculated. The macro-averaged metrics, which treat all hazard levels equally, indicated a precision of 0.527, recall of 0.471, and a macro F1-score of 0.487. These values suggest that the model was moderately effective across all classes but struggled particularly with minority hazard categories. By contrast, the weighted averages, which account for class prevalence, were higher: precision = 0.598, recall = 0.612, and F1 = 0.602. This discrepancy highlights the model’s tendency to favour performance on the more common Low and Moderate hazard levels while underperforming on rarer categories such as Considerable+ and High.

A critical indicator of classification reliability in multiclass problems is Cohen’s Kappa coefficient, which adjusts for agreement by chance. The model achieved a Kappa score of 0.449, placing it in the “moderate agreement” range according to standard interpretation thresholds. This confirms that the network’s predictions contain meaningful signal beyond random guessing, but that performance remains far from perfect reliability. The EDA findings provide context for this outcome: high missingness in some snowpack variables (>20%) required their removal, potentially reducing predictive richness, while outlier analysis showed extreme weather and snow values that could distort learning, especially in minority hazard levels.

Given the ordinal nature of avalanche hazard levels, error magnitude is as important as classification accuracy. The model achieved a Mean Absolute Error (MAE) of 0.450 hazard levels on the 0–4 scale, indicating that predictions were on average less than half a category away from the true hazard level. Similarly, the Root Mean Squared Error (RMSE) was 0.765, and the Mean Squared Error (MSE) was 0.585, reflecting relatively low dispersion in error magnitudes. These values confirm that, even when misclassifications occurred, they were usually small in ordinal distance. This aligns with the EDA’s feature relationship analysis, which showed that many predictors (e.g., snow temperature, drift, precipitation type) were moderately correlated with hazard levels. The network was able to capture these associations, but the overlap between Moderate and Considerable- categories made fine-grained separation difficult, leading to smaller but frequent ordinal misclassifications.

This ordinal performance is further supported by adjacent accuracy measures. The model achieved an Adjacent Accuracy of 94.3%, meaning that nearly all predictions were within ±1 hazard level of the correct category. The Within-2 Accuracy was 99.5%, and Within-3 Accuracy reached 100%, demonstrating that large errors (e.g., predicting Low when the actual level was High) were extremely rare. This is highly consistent with the EDA summary, which emphasized that extreme misclassifications were unlikely because strong meteorological signals (e.g., heavy snowfall or very high wind speeds) rarely coincided with Low hazard levels, giving the model clear boundaries for the extremes.

Ordinal correlation metrics confirmed that the network captured the underlying ordered structure of the hazard levels. The Spearman’s Rank Correlation coefficient was 0.736, while the Pearson correlation was 0.720, both indicating strong monotonic associations between predicted and actual hazard categories. Additionally, Kendall’s Tau was 0.669 (p < 0.001), reinforcing that the network preserved risk ordering with high fidelity. Finally, directional accuracy, which evaluates whether the model correctly identified the trend in risk comparisons between instances, was 0.714, showing that the majority of pairwise hazard comparisons were ordered correctly.

Collectively, these results demonstrate that while the network achieved only moderate exact-match classification accuracy, its ordinal-aware performance metrics paint a more favourable picture. The model rarely produced large misclassifications, maintained a strong correlation with true hazard orderings, and delivered substantial improvements over naive baselines. The integration of EDA insights explains why performance was concentrated around the common hazard levels and why small misclassifications were frequent, but catastrophic errors were avoided. This combination of performance characteristics suggests that the neural network has practical utility in avalanche forecasting, particularly when supported by confidence-based uncertainty estimates and operational interpretation.

**Confusion Matrix and Class-Level Performance**

The confusion matrix provided a detailed view of the network’s predictive behaviour across the five hazard levels. The results confirmed the overall trend observed in the weighted metrics: strong predictive ability for majority classes (Low and Moderate), moderate reliability for intermediate conditions (Considerable–), and poor performance on rarer high-hazard categories (Considerable+ and High).

For the **Low hazard class**, the model achieved a sensitivity of **0.798** and precision of **0.734**, indicating consistent identification of stable snowpack conditions. This aligns with EDA findings that Low hazard days were characterized by clear predictor signatures — shallow snow depth, low drift, and stable temperature gradients — which the model could learn effectively. The **Moderate class** followed a similar trend, with sensitivity of **0.597** and precision of **0.562**, reflecting reliable but imperfect detection of these common conditions. Together, these two classes accounted for the majority of correct predictions, mirroring their dominance in the dataset distribution.

The **Considerable– class** achieved moderate scores (precision = **0.554**, recall = **0.541**). Errors were frequent between adjacent categories, particularly with Moderate (87 misclassifications) and Considerable+ (17 misclassifications). This reflects the transitional nature of the predictors observed in the EDA, where snow depth, crystal type, and summit air temperature exhibited overlapping ranges between Moderate and Considerable– conditions.

Performance dropped sharply for **Considerable+** and **High hazard** levels. Considerable+ achieved precision of **0.347** and recall of **0.205**, while High hazard registered precision of **0.438** and recall of **0.212**. These results indicate systematic under-detection of severe instability, with most High cases misclassified as Considerable– (13) or Considerable+ (11). Importantly, there were no cases of High being misclassified as Low, confirming that large ordinal errors were avoided. Instead, misclassifications clustered around neighbouring categories, consistent with the strong ordinal correlations observed in the overall metrics.

The **per-class F1-scores** further illustrated the gradient in performance: Low (**0.765**) and Moderate (**0.579**) achieved acceptable balance between precision and recall, while Considerable– was moderate (**0.548**). In contrast, Considerable+ (**0.258**) and High (**0.286**) had very low F1-scores, highlighting their vulnerability to both false negatives and false positives. This weakness reflects the class imbalance noted in the EDA, where these categories made up less than 5% of the dataset, and the overlapping meteorological signals that blur distinctions between upper hazard levels (e.g., heavy snowfall, wind drift, and deep weak layers).

Overall, the confusion matrix underscores a key strength and weakness of the neural network. The model excelled in predicting stable and moderately unstable conditions, which dominate the operational forecast space, but it lacked sufficient discriminative power for the most critical high-risk categories. This imbalance in predictive skill raises important implications for operational use: while the network can be trusted to support routine decision-making in low-to-moderate risk contexts, its outputs for rare but high-consequence events require cautious interpretation and likely supplementation with expert judgment or additional modelling approaches.

**Confidence-Based Reliability**

To further assess the reliability of the neural network’s outputs, prediction probabilities were stratified into confidence levels. Out of the 1,128 test samples, 122 predictions (10.8%) were made with high confidence (>0.8), 331 (29.3%) with medium confidence (0.6–0.8), and the majority, 675 (59.8%), with low confidence (≤0.6). This distribution indicates that the network was generally conservative in assigning strong certainty to its predictions, reflecting the inherent difficulty of separating avalanche hazard categories in complex meteorological settings.

Accuracy strongly depended on prediction confidence. High-confidence outputs achieved an accuracy of 90.2%, demonstrating that when the network was certain, its predictions were highly reliable. Medium-confidence predictions achieved 71.0% accuracy, while low-confidence predictions dropped to 51.1%, only marginally above random chance. This gradient illustrates that prediction probabilities served as meaningful proxies for reliability, allowing forecasters to weigh model outputs according to their associated confidence.

These patterns were consistent with the EDA findings. Variables such as snow depth, wind speed, and temperature gradients displayed substantial variability within intermediate hazard categories (Moderate and Considerable–), leading to overlapping predictor distributions. This overlap likely contributed to the large proportion of low-confidence predictions, as the model encountered ambiguous feature configurations where class boundaries were not clearly defined. Conversely, high-confidence predictions were more common in the Low hazard class, which the EDA showed to have stable and distinctive predictor signatures (e.g., shallow snow depth, minimal drift, and stable thermal gradients).

From an operational perspective, these results suggest that incorporating prediction confidence could substantially enhance the model’s utility. High-confidence predictions, which reached over 90% accuracy, could be integrated directly into forecasting workflows, while low-confidence outputs should be flagged for additional human review or supplementary modelling.

Thus, while the model’s overall accuracy was moderate, the stratification of performance by confidence level provides a valuable operational safeguard. Confidence-based filtering allows practitioners to trust high-certainty predictions while exercising caution in ambiguous cases, thereby supporting safer and more robust integration of machine learning into avalanche risk assessment.

**Avalanche-Specific Safety Metrics**

The model’s safety evaluation revealed both conservative tendencies and critical weaknesses. The Critical Miss Rate was 21.6%, meaning over one in five severe avalanche cases (Considerable+ or High) were underestimated as Low or Moderate. This was linked to overlapping predictor patterns between Moderate and higher levels, as identified in the EDA, and the scarcity of severe events (<5% of cases).

Despite this, the model achieved 74.9% Safety Effectiveness, correctly predicting at or above the true hazard level in most cases. However, the Conservative Bias was slightly negative (–0.114), reflecting a mild overall tendency to underestimate risk. High-risk detection showed mixed results: Sensitivity was 67.9%, indicating some missed severe cases, while Specificity reached 91.8%, showing reliable performance in avoiding false alarms.

**Feature Importance Assessment**

The multi-method feature selection process revealed a clear hierarchy of predictors, many of which aligned with established avalanche science and the exploratory findings. Foot Penetration emerged as the most influential feature, reflecting its role as a proxy for snowpack weakness and load-bearing capacity. This was followed by Drift, which is consistent with the importance of wind transport in forming unstable slabs. Summit Air Temperature and Total Snow Depth were also highly ranked, capturing atmospheric drivers and snowpack accumulation processes that the EDA had shown to vary significantly across hazard levels. The inclusion of Crystal Type among the top predictors confirmed the influence of snow microstructure, particularly in transitions from Moderate to Considerable- hazard conditions.

Thermal and meteorological interactions were also critical. Variables such as Air Temperature, Wind Chill, and Wind Speed consistently scored highly across LASSO, Random Forest, and correlation methods. Their prominence reflects the operationally recognised role of strong winds and steep temperature gradients in weakening snow layers and triggering instability, The observed correlations between summit and base temperatures during EDA further supported the predictive role of atmospheric gradients.

Temporal and seasonal features carried significant predictive weight as well. The Winter season, Day of Year, and Snow–Altitude Interaction ranked among the strongest predictors outside the direct snowpack variables. The EDA confirmed these seasonal structures, showing distinct peaks in hazard ratings during winter months and at higher altitudes.

Geographical indicators, including Area (Lochaber and Torridon), were also retained in the final set of 25 features. While their predictive importance was secondary to meteorological and snowpack measurements, their inclusion reflects the spatial heterogeneity of avalanche processes in Scotland. This is consistent with both the EDA results and regional avalanche research showing that terrain-modulated differences (e.g., exposure, wind redistribution) influence local hazard distributions

Together, these results confirm that the model’s strongest predictors aligned with well-established drivers of avalanche instability: snowpack structure, depth, wind transport, temperature gradients, and seasonality.

A group of graphs and diagrams

AI-generated content may be incorrect.

**Conclusion and recommendations**

The neural network achieved 61.2% accuracy on the test set, an 81.6% improvement over the baseline. While exact classification was moderate, ordinal metrics showed strong performance, with 94.3% of predictions within ±1 hazard level and a mean absolute error of 0.45 levels. Large misclassifications were rare, and the model preserved the ordered structure of avalanche danger ratings. Performance was strongest for Low and Moderate hazards but weaker for Considerable+ and High, reflecting class imbalance noted in the EDA. The critical miss rate of 21.6% highlights risks of underestimating severe conditions, though the model generally erred conservatively, with 74.9% safety effectiveness. Key predictors included Foot Penetration, Drift, Summit Air Temperature, Total Snow Depth, and Crystal structure, alongside temporal and seasonal variables, confirming both snowpack and climatic influences on avalanche hazard.

To improve performance, future work should focus on better representation of high-hazard cases, possibly through expanded datasets or transfer learning. Ensemble methods and uncertainty-based outputs could also increase operational reliability. While not yet suitable as a standalone system, the model shows strong potential as a decision-support tool when combined with expert judgment.

**References (APA)**

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

European Avalanche Warning Services. (n.d.). *Danger scale and standards*. [https://www.avalanches.org](https://www.avalanches.org/)

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Juvik, E. S. (2023). Toward data-driven decision support for avalanche-exposed roads in Norway. In *Proceedings of the International Snow Science Workshop* (ISSW 2023). Norwegian Public Roads Administration.

McClung, D., & Schaerer, P. (2006). *The avalanche handbook* (3rd ed.). The Mountaineers Books.

Scottish Avalanche Information Service. (n.d.). *About SAIS and daily avalanche reports*. [https://www.sais.gov.uk](https://www.sais.gov.uk/?utm_source=chatgpt.com)

Schweizer, J., Jamieson, B., & Schneebeli, M. (2003). Snow avalanche formation. *Reviews of Geophysics, 41*(4), 1–25. <https://doi.org/10.1029/2002RG000123>

Techel, F., & Zweifel, B. (2014). Recreational avalanche accidents in Switzerland: Trends and patterns with an emphasis on burial, rescue methods and avalanche danger. *Cold Regions Science and Technology, 97*, 77–85. <https://doi.org/10.1016/j.coldregions.2013.09.006>

Techel, F., Jarry, F., Kronthaler, G., Mitterer, C., Nairz, P., Pavšek, M., Valt, M., & Zweifel, B. (2016). Avalanche fatalities in the European Alps: Long-term trends and perspectives. *Cold Regions Science and Technology, 121*, 22–31. <https://doi.org/10.1016/j.coldregions.2015.11.017>