**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 comparison of a graph

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**

#DATA PREPROCESSING

After initial data cleaning, we removed non-predictive fields (OAH, Obs) and filtered out the 1% of rows with missing FAH which left us with 10,562 usable records.The FAH values were standardized into five clean categories with the following counts: Low 3,433, Moderate 3,239, Considerable− 2,498, Considerable+ 933, High 459 this was confirming a strong skew toward safer days.

A missingness audit showed two variables with >20% gaps—AV.Cat (23.4%) and Ski.Pen (22.5%) these we then dropped to avoid heavy, low-confidence imputation. The next tier of missingness is moderate and structured rather than random. Core near-surface weather and geometry variables have very low gaps (≤1–3%). These patterns motivate our Area×Season imputation (fit on the training split only).

# OUTLIER DETECTION AND REMOVAL

To stabilise modelling while preserving the signal, we applied a two-tier outlier process. First, we converted any string “NA”s to true NAs and used a robust IQR filter (k = 3.5) on key continuous variables, while looking to prune only extreme spikes: Alt (2), Wind.Speed (22), Summit.Wind.Speed (81), Total.Snow.Depth (111), Air.Temp (1), Summit.Air.Temp (0), and Foot.Pen (34). Second, we enforced physical plausibility rules and essentially removed Aspect > 360° (8), constrained Cloud to [0,100]% (3), and limited Incline to [0,90]° (5). In total, 267 rows (2.53%) were dropped. This keeps legitimate extremes (e.g., high winds/depths during storms) while eliminating impossible or clearly erroneous readings that would distort scaling, inflate variance, and hinder neural-network optimisation.

# FEATURE ENGINEERING

We developed a small set of mechanism-aware feature**s** to turn raw measurements into signals the model can learn from. A *Wind\_Chill* proxy (*Air.Temp − 0.6×Wind.Speed*) captures near-surface cooling that weakens bonds under strong winds. A *Temp\_Gradient* (*Summit.Air.Temp − Air.Temp*) approximates vertical stability and transport potential. *Snow\_Alt\_Interaction* (*Total.Snow.Depth × Alt/1000*) lets snowfall load scale with elevation and the rain–snow line. Because *Aspect* is circular (0°≈360°), we encoded it as *Aspect\_North* **=** *cos(Aspect)* and *Aspect\_East* **=** *sin(Aspect)* to avoid artificial discontinuities and to align with wind-loading geometry. Finally, we added *Month*, *Day\_of\_Year*, and *Season* to reflect the strong winter seasonality seen in the EDA and to let the network learn intra-season cycles (e.g., cold spells vs. thaw pulses). Collectively, these features reduce redundancy, respect data geometry, and embed avalanche mechanics directly into the predictors.

# IMPUTATION

We address missing data with a context-aware, hierarchical imputation routine that mirrors how avalanche conditions vary across space and season. First, for each Area×Season group we fill numeric gaps with thegroup median(robust to outliers) and categorical gaps with the group mode, so replacements reflect typical conditions for that region and time of year rather than a blunt global average. Next, any leftovers are handled with global medians/modes as a conservative fallback. We explicitly exclude non-predictive and target fields (Date, OSgrid, Location, FAH) from imputation to avoid leakage into the label. This design preserves regional/seasonal structure noted in the EDA while producing a complete feature matrix for modelling.

# TARGET ENCODING

To prepare the data for a neural network, we first removed non-predictive IDs (Date, OSgrid, Location) and encoded the target, FAH, as an ordinal integer in true risk orderLow→High = 0–4. A verification table confirmed a one-to-one mapping (Low=0, Moderate=1, Considerable−=2, Considerable+=3, High=4) with no unmapped values and the encoded class counts **[**3398, 3167, 2412, 895, 423], matching the observed imbalance.

For predictors, we cleaned categorical fields and applied a frequency threshold (≥50) so very rare levels were folded into “Other” to limit sparsity. We then one-hot encoded the remaining categories for Area, Precip.Code, and Season, while adding binary indicator columns and dropping the original fields. This pipeline preserves the ordinal meaning of the hazard label while transforming categorical inputs into a numerically stable representation that dense neural networks can learn from, with reduced risk of overfitting to tiny categories.

# FEATURE SELECTION

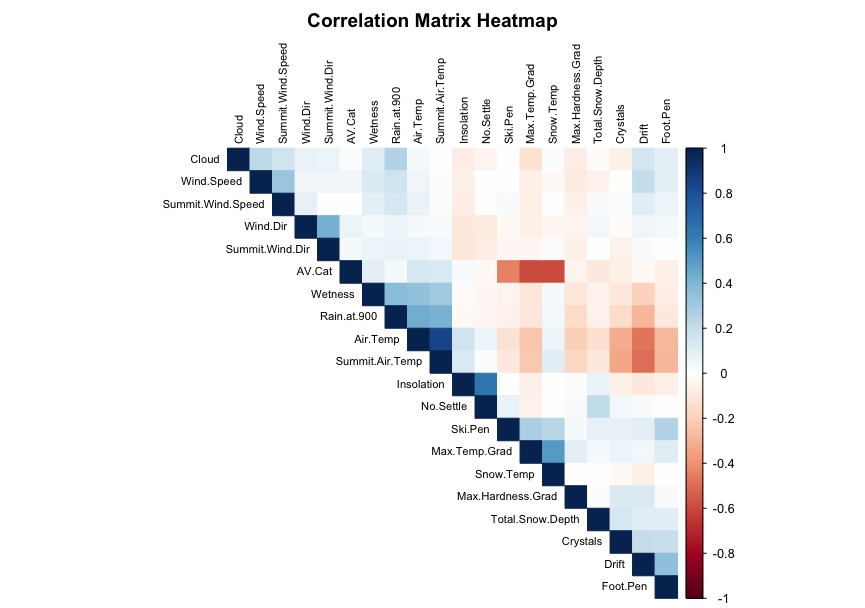
After cleaning and encoding, we standardized all numeric predictors and ran an ensemble feature-selection procedure on the 7,530 complete cases to concentrate signal and control dimensionality. Three complementary signals were computed per feature—LASSO coefficients (linear sparsity), Random Forest permutation importance (non-linear/interaction effects), and target correlation—each normalized to 0–100 and combined with weights (RF 40%, LASSO 30%, Correlation 30%). LASSO retained 30 features, Random Forest and correlation evaluated 47 while the weighted scores yielded a final shortlist of 25 predictors. The top-ranked variables are physically coherent with avalanche mechanics are

* *Foot.Pen* (near-surface strength)
* *Drift* (recent wind transport)
* *Summit.Air.Temp* and *Air.Temp* (thermal state)
* *Total.Snow.Depth* (load)
* *Wind.Speed/Summit.Wind.Speed* (loading potential)
* *Max.Temp.Grad* (faceting potential)
* *Snow.Temp*
* *Insolation*
* *Wind\_Chill, Snow\_Alt\_Interaction,* and *Day\_of\_Year* (engineered terms)
* *Precip.Code*
* *Season\_Winter*
* *Area\_Lochaber* and *Area\_Torridon*

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.

**Subsection: Feature engineering**



The correlation heatmap in Figure X above reveals coherent but mostly moderate correlation clusters among variables. Notably, temperatures (Air.Temp, Summit.Air.Temp) exhibit strong covariation, while Cloud and Insolation show the expected inverse relationship. Wind speeds correlate across levels with some directional noise, and penetration/strength metrics (Foot.Pen, Ski.Pen) associate with Total.Snow.Depth in predictable ways. Meanwhile, Max.Temp.Grad and Max.Hardness.Grad relate to temperature and snow-structure variables through more complex mechanisms, likely tied to metamorphic processes in the snowpack. Overall, this indicates potential multicollinearity, particularly among thermodynamic and wind features, which could destabilize linear models and hinder neural network convergence. To address this, we implemented ensemble feature selection combining LASSO regularization, random forest importance scoring, and correlation-based filtering to minimize redundancy.

Leveraging these patterns, we engineered targeted features to capture key avalanche mechanics in stable, low-redundancy forms:

* **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*) as a 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 dataset of 7,530 observations was split into training (70%), validation (15%), and test (15%) sets, stratified by hazard level to maintain class proportions. This yielded 5,273 training samples, 1,129 validation samples, and 1,128 test samples from the 7,530 complete cases. The training set reflected the dataset's imbalance, with Low (class 0) and Moderate (class 1) each >1,700 samples, versus only 145 for High (class 4)—an imbalance ratio of ~12.4:1. Validation and test sets preserved similar distributions, retaining the rarity of high-hazard classes.

To mitigate bias toward majority classes, ROSE balancing was applied solely to the training data post-split, keeping validation/test sets natural for unbiased evaluation. Using an iterative one-versus-rest approach (since ROSE is binary-native), synthetic samples were generated for minorities until they reached ~95% of majority size, avoiding perfect balance.

The training set grew from 5,273 to 8,641 samples, with ~1,700 per class, reducing the imbalance ratio to 1.05:1 (a 91.5% improvement). This enhanced minority representation for neural network training while ensuring realistic 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 neural network’s performance was assessed across the training, validation, and test datasets to evaluate both its learning capacity and generalization ability. During training, the model achieved a best validation accuracy of 60.2%, closely matching its final test accuracy of 61.2% (loss = 0.978). This alignment between validation and test performance suggests that the model did not overfit during training and generalized reasonably well to unseen data. By contrast, the naive baseline accuracy, derived from the most frequent class (Moderate, 33.7% of the test set), was substantially lower, meaning the network delivered an 81.6% relative improvement over this trivial predictor.

The exploratory data analysis (EDA) helps contextualize this outcome. The FAH distribution was highly imbalanced, with Low and Moderate levels dominating and High being rare. Although the training set was balanced using ROSE, the validation and test sets retained the natural skew, which explains why the model consistently performed better on common classes than on rare ones. Geographical and seasonal patterns also shaped this performance: areas like Lochaber and Torridon contributed disproportionately more samples, while High hazard levels clustered during winter peaks. These structural imbalances likely reinforced the network’s ability to learn Low and Moderate risks while limiting its accuracy on rarer categories.

Beyond accuracy, complementary metrics provided a broader view of generalization. The macro-averaged scores (precision = 0.527, recall = 0.471, F1 = 0.487) showed that the model was only moderately effective when all classes were weighted equally. However, the weighted averages (precision = 0.598, recall = 0.612, F1 = 0.602) were notably higher, reflecting stronger performance on the dominant classes. Cohen’s Kappa, at 0.449, placed the model in the “moderate agreement” range, confirming that predictions were significantly better than random guessing but not yet highly reliable. EDA findings again provide insight: the removal of several snowpack variables with >20% missingness reduced predictive richness, while the presence of extreme but plausible outliers (e.g., high winds, deep snow) may have complicated learning for rarer hazard levels.

Given the ordinal nature of avalanche hazards, the model’s error magnitudes were as critical as accuracy. Predictions deviated by an average of only 0.45 hazard levels (MAE), with relatively low dispersion (RMSE = 0.765). Adjacent accuracy was exceptionally high (94.3% within ±1 level), while extreme errors were nearly absent (99.5% within ±2 levels; 100% within ±3 levels). These results confirm that large misclassifications, such as predicting Low when the true level was High, were virtually eliminated. This aligns with EDA findings showing that extreme meteorological conditions—like heavy snowfall or very high winds—rarely coincided with Low hazard levels, giving the model strong boundaries at the extremes.

Finally, correlation-based measures demonstrated the network’s ability to preserve ordinal structure. Strong monotonic relationships were observed between predicted and actual hazard levels (Spearman’s ρ = 0.736; Pearson r = 0.720), while Kendall’s Tau (0.669, p < 0.001) and directional accuracy (0.714) indicated consistent ordering across hazard categories. Taken together, these results reveal that although the model achieved only moderate exact-match accuracy, it maintained ordinal consistency, avoided catastrophic misclassifications, and delivered meaningful improvements over baseline predictors.

**Confusion Matrix and Class-Level Performance**

The confusion matrix offered a clear breakdown of how the model performed across individual hazard levels. For the Low hazard class, the network achieved strong results (sensitivity = 0.798, precision = 0.734), reflecting consistent identification of stable snowpack conditions. This strength is consistent with EDA findings, which showed that Low days were associated with distinct signals — shallow snow depth, weak drift, and stable temperature gradients — making them easier for the model to separate. Similarly, the Moderate class reached moderate reliability (sensitivity = 0.597, precision = 0.562), supported by its high prevalence in the dataset and relatively consistent meteorological profiles. Together, these two classes accounted for most correct predictions, reflecting both their dominance in the dataset and their clearer predictor signatures.

The Considerable– class performed less consistently (precision = 0.554, recall = 0.541). Misclassifications were concentrated between adjacent levels: 87 cases confused with Moderate and 17 with Considerable+. This behaviour aligns with the transitional nature of predictors highlighted in the EDA, where snow depth, crystal type, and summit air temperature overlapped substantially between Moderate and Considerable– categories.

Performance deteriorated for the Considerable+ and High hazard levels, which achieved very low recall (0.205 and 0.212, respectively). High cases were most often misclassified as Considerable– (13) or Considerable+ (11), but crucially, none were mistaken for Low. This indicates that while the model struggled to distinguish among upper hazard levels, it preserved ordinal structure and avoided catastrophic misclassifications. The EDA findings explain this limitation: these rare categories comprised less than 5% of the dataset and were characterized by overlapping meteorological signals such as heavy snowfall, wind drift, and deep weak layers. With limited training samples, the network could not fully disentangle these patterns.

The per-class F1-scores reflected this gradient in performance: Low (0.765) and Moderate (0.579) were acceptable, Considerable– was moderate (0.548), while Considerable+ (0.258) and High (0.286) were poor. These values highlight the central trade-off of the model: strong utility in predicting stable and moderately unstable conditions, but weak reliability in capturing rare but operationally critical high-risk categories.

**Confidence-Based Reliability**

The analysis of prediction probabilities revealed that most test predictions were made with low confidence (≤0.6, 59.8%), while only a small share were high-confidence (>0.8, 10.8%). This indicates that the model was generally conservative in assigning strong certainty, reflecting the inherent difficulty of distinguishing avalanche hazard levels under overlapping meteorological conditions.

Accuracy closely followed confidence level. High-confidence predictions achieved 90.2% accuracy, showing that when the network was certain, it was usually correct. Medium-confidence predictions (71.0%) were moderately reliable, while low-confidence predictions (51.1%) approached random chance. This gradient demonstrates that prediction probabilities serve as a valid proxy for model reliability.

These findings are consistent with the EDA results. The overlap of predictor distributions in intermediate hazard levels especially Moderate and Considerable– produced ambiguous cases that drove the majority of low-confidence predictions. By contrast, high-confidence predictions were concentrated in the Low hazard class, where EDA showed distinctive conditions such as shallow snow depth, minimal drift, and stable thermal gradients.

Operationally, these results highlight the importance of confidence-aware interpretation. High-confidence predictions could be used directly in forecasting workflows, while low-confidence outputs should be flagged for expert review or supplemented by additional modelling. This stratified reliability provides an operational safeguard, ensuring that machine learning support enhances decision-making without introducing undue risk in ambiguous cases.

**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**

Th The feature importance analysis revealed that predictors strongly aligned with established avalanche science and EDA findings. Foot Penetration and Drift were the top-ranked features, reflecting their direct links to snowpack weakness and wind-driven slab formation. Key atmospheric drivers — Summit Air Temperature and Total Snow Depth — also ranked highly, confirming their influence on hazard variability observed in the EDA. Crystal Type added further insight into snow microstructure transitions, especially between Moderate and Considerable– categories.

Wind- and temperature-related factors such as Air Temperature, Wind Chill, and Wind Speed consistently scored highly across methods, echoing EDA findings on strong thermal gradients and wind redistribution as critical instability drivers. Seasonal and temporal features — particularly Winter, Day of Year, and Snow–Altitude Interaction — captured the clustering of high hazard levels during winter peaks and at higher elevations. Geographical indicators such as Lochaber and Torridon added spatial context, though with secondary importance.

In summary, the most important features reflected a combination of snowpack properties, meteorological drivers, and seasonal patterns, confirming that the model’s predictive behaviour was rooted in meaningful physical processes identified during the exploratory analysis.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>