Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking Genomic Breed Discrimination in Cattle and Sheep: Discrimination, Calibration, and the Role of Population Structure

Yalçın YAMAN

yalcinyaman@gmail.com

Siirt Üniversitesi

Burcu TOKGÖZ

Siirt Üniversitesi

Article

Keywords: Genomic Breed Discrimination, Machine Learning in Livestock, Breed-Informative SNPs, Autoencoder Dimensionality Reduction, Livestock Traceability

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10527337/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Benchmarking Genomic Breed Discrimination in Cattle and Sheep: Discrimination, Calibration, and the Role of Population Structure

Yalçın YAMAN[1*], Burcu TOKGÖZ[1]

[1]*Faculty of Veterinary Medicine, Siirt University, TR-56100, Siirt – TÜRKİYE*

*** Correspondence:**
Yalçın YAMAN: yalcinyaman@gmail.com
yalcin.yaman@siirt.edu.tr

Abstract

Background. Reliable genomic breed assignment underpins livestock traceability, pedigree verification, and conservation breeding, yet most benchmarks address a single species and rank classifiers by discrimination alone.

Methods. We benchmarked fourteen machine-learning and statistical classifiers for breed assignment in 20 cattle breeds (n = 403; *Bos taurus*, *B. indicus*, and *B. javanicus* lineages) and 20 *Ovis aries* sheep breeds (n = 1,458). Whole-chip SNP data underwent quality control, LD pruning, and autoencoder-based feature selection to 5,000 informative markers, and models were compared across discrimination, calibration, and generalisation using a composite weighted multi-metric ranking.

Results. Top classifiers reached near-maximal accuracy in both species, but inter-model structure differed. In cattle, linear and probabilistic models — with both SVM variants — formed a statistically indistinguishable top tier, whereas boosting models showed train–test gaps consistent with overfitting under the high-dimensional, small-sample design. In sheep, inter-model variance was markedly compressed, consistent with the more homogeneous *Ovis aries* structure. A systematic accuracy–calibration dissociation emerged in both species: ridge regression and KRR (a GBLUP-analogous KernelRidge baseline with softmax-normalised, non-probabilistic posteriors) discriminated well but were poorly calibrated. PyTorch MLP, logistic regression, and elastic net occupied the top three positions, combining high accuracy with comparatively reliable posteriors. Residual misassignment — Braunvieh in cattle and five

commercially selected European-derived sheep breeds — was shared across algorithmically distinct classifiers, indicating population-level genomic proximity rather than algorithmic limitation.

Conclusions. Accurate, well-calibrated breed assignment is attainable in both species, and classification difficulty tracked population structure more than algorithm choice. As findings derive from within-sample cross-validation, extension to independent or admixed animals requires external validation.

Keywords: Genomic Breed Discrimination; Machine Learning in Livestock; Breed-Informative SNPs; Autoencoder Dimensionality Reduction; Livestock Traceability

1. Introduction

The reliable identification of livestock breeds constitutes a foundational requirement of modern animal production, underpinning the conservation of locally adapted genetic resources, the traceability of animal-derived food products, the integrity of international breeding markets, and the validity of genotype–phenotype association studies [1,2]. As genomic selection programmes expand globally, unambiguous assignment of animals to their breed of origin has become progressively more important: errors in breed records propagate into estimated breeding values, inflate residual variance in mixed-model evaluations, and compromise authenticity guarantees in premium supply chains. Historically, breed identification relied on pedigree documentation and morphological evaluation. Pedigree-based systems are susceptible to recording errors and genealogical gaps, particularly in smallholder and pastoral settings where formal registries are absent or inconsistently maintained [1]. Morphological assessment is often constrained by the phenotypic plasticity of traits such as coat colour, horn morphology, and body conformation, which can respond to nutritional status, age, and environmental conditions in ways that may be difficult to disentangle from breed-level variation. These challenges are further compounded when discriminating between morphologically proximate breeds — such as those sharing recent common ancestry or subject to parallel selection pressures — where phenotypic overlap can render visual assignment unreliable even under expert evaluation. Subjectivity in scoring protocols may additionally limit the reproducibility and cross-study comparability of phenotype-based assignment, and together these

considerations have contributed to growing interest in objective, high-throughput molecular approaches as complementary or alternative tools for breed identification.

Commercial high-density single nucleotide polymorphism (SNP) genotyping arrays have catalysed a progressive shift towards molecular breed identification. Genome-wide SNP panels encode patterns of linkage disequilibrium, population stratification, and breed-specific selective sweeps that together provide what may be regarded as a breed's genomic signature — a heritable signal that is largely stable across environments and directly quantifiable, subject to genotyping error [2,3]. The standard analytical pipeline — marker quality control, selection of breed-informative SNPs, and supervised classification — has been systematically reviewed, with recent recommendations favouring selection procedures that balance discriminatory power and computational tractability [1,3]. The choice of biSNP selection criterion, whether based on allele frequency divergence metrics such as F_{st} and Delta or on multivariate projections such as partial least squares regression, materially influences downstream classification accuracy, underscoring the importance of systematic cross-species benchmarking. Although whole-genome sequence data tend to yield marginally higher identification accuracy than medium-density chip data, the practical difference is frequently modest, suggesting that 50K–100K SNP arrays represent a pragmatic compromise for most operational breeding programmes [2].

The application of machine learning (ML) to livestock breed discrimination has a trajectory spanning more than fifteen years. Early work by Burocziova and Riha[4] demonstrated that probabilistic classifiers — including Naïve Bayes, Bayes Net, and instance-based methods — could reliably discriminate among eight horse breeds from microsatellite data, establishing a methodological precedent for applying artificial intelligence to population-level breed assignment. The subsequent availability of high-density SNP arrays substantially expanded the discriminatory power of such approaches. Traditional parametric discriminants — including linear discriminant analysis and principal component regression — have been argued to be limited in capturing non-linear or epistatic structure where this contributes to breed-specific genetic architecture [5]. ML classifiers impose minimal distributional assumptions while accommodating the high-dimensional, correlated structure of SNP data. The extent to which breed discrimination depends on non-linear, as opposed to additive allele-frequency, signal is, however, dataset-dependent and remains an open question — one that a systematic comparison of linear, regularised, ensemble, and deep-learning classifiers is well placed to address.

Individual algorithms such as support vector machines and random forests have been reported to achieve identification accuracies of 95% or above in well-structured reference populations, and ensemble integration of complementary breed-informative SNP selectors and classifiers consistently yields higher performance than any single-method approach [2]. A comparative study across pig breeds demonstrated that no single algorithm uniformly dominates: different classification approaches offer complementary strengths depending on the number of discriminant markers and the degree of population stratification [6].

More recent work has extended these approaches using deep learning architectures. Manzoori et al.[7] applied deep neural networks alongside Garson and Olden saliency methods to identify discriminant SNP panels across 37 horse breeds, demonstrating that deep learning-derived feature selection achieves population assignment with substantially fewer markers than conventional approaches. Multi-Layer Perceptron architectures have extended performance further, with a multi-output regression MLP achieving full discrimination of purebred and hybrid individuals across multiple SNP chip densities in pigs [8]. Ensemble and stacking strategies have similarly demonstrated robust predictive performance across diverse pig genomic evaluation tasks, ranging from breed-specific identification [6] to complex phenotypic prediction [9]. The translational applicability of these frameworks is illustrated by web-accessible tools such as iDIGs, which classifies 91 pig breeds from as few as 3,000 SNPs at 98.4% cross-validated accuracy [10], and by rPIMS, a ShinyR application that integrates PCA-based dimensionality reduction, phylogenetic tree construction, population structure analysis, and multiple ML classifiers within a single graphical interface, reporting 100% classification accuracy across 10 livestock breeds using 860 SNPs. Notably, the developers of rPIMS explicitly recognised that previous breed discrimination studies have predominantly focused on single species or commercially selected multi-breed panels, often requiring large training datasets not readily available for local or indigenous populations [11].

A persistent methodological challenge concerns performance under admixture. When individuals carry partial genomic contributions from multiple breeds, classification error rates rise markedly; in cattle, misclassification rates of up to 7.82% have been reported for historically admixed breeds against a population-wide mean of approximately 1.55% [2]. Modelling genomic breed composition as a continuous regression output — regressing individual SNP genotypes on reference breed allele frequencies — offers a principled approach, simultaneously estimating purebred probability, quantifying breed proportions in

crossbred individuals, and facilitating cross-breed genetic evaluations [2,3]. A comparison of SNP-BLUP and Admixture across thirteen cattle breeds demonstrated that both approaches perform comparably when at least 3,000 informative SNPs are considered, and that breed composition prediction may be integrated into existing genomic evaluation pipelines without dedicated standalone software [12]. The genomic relationship matrix (GRM) has also been proposed as an alternative classifier input that bypasses the explicit breed-informative SNP selection step, providing robust breed assignment in populations with moderate differentiation [3].

Despite the substantial progress summarised above, several methodological gaps remain unaddressed. First, existing breed discrimination frameworks have been developed and validated within individual species — a limitation explicitly recognised in the literature, as previous studies have mostly focused on single species or commercially selected multi-breed populations [11]. Systematic benchmarking of the same analytical pipeline across two livestock species, represented by independent reference datasets, has not previously been reported; such a comparison allows classifier behaviour to be examined under contrasting dataset structures. Second, the number of classifiers evaluated in any single study has typically been limited to three or four algorithms, spanning four model families (parametric, regularised, ensemble, and deep learning) [11,13,14]. Third, autoencoder-based dimensionality reduction — which learns a non-linear compact representation of genome-wide SNP variation in a data-driven manner — has rarely, if ever, been applied as a pre-classification feature selection step in breed discrimination workflows. Existing studies rely principally on linear projections such as PCA or on supervised saliency scores derived from trained classifiers [3,11,15]. Fourth, and perhaps most importantly, breed panels in the literature typically encompass breeds from a single continent, subspecies, or production system. A panel spanning taurine European and American breeds, humped *Bos indicus* breeds, and *Bos javanicus* alongside sheep populations from Europe, Africa, the Middle East, the Americas, and Central Asia presents substantially higher levels of both within-group genetic heterogeneity and between-group divergence, providing a broad and heterogeneous evaluation setting for classification algorithms.

The present study addresses these gaps through a systematic benchmarking of SNP-based genomic breed discrimination across 20 cattle breeds (n = 403) — comprising taurine European breeds (Angus, Aubrac, Braunvieh, Gascon, Holstein, Maraichine, Normande, Piedmontese, Romagnola, Salers, Shorthorn), indicine breeds (Gir, Ongole, Sahiwal, Zebu_general), African

breeds (Baoule, Lagune, N'Dama, Oulmes Zaer), and the domesticated Bali cattle — and 20 sheep breeds (n = 1,458) spanning European (Australian Poll Dorset, Australian Suffolk, Border Leicester, Churra, Finnsheep, Galway, German Texel, Irish Suffolk, Meat Lacaune, Milk Lacaune, Rambouillet, Scottish Blackface, Scottish Texel, Soay), American (Gulf Coast Native, Santa Ines), African (Red Maasai), Middle Eastern (Afshari, Qezel), and Central Asian (Tibetan) populations. Prior to classification, an autoencoder deep learning approach was applied to genome-wide SNP data to identify informationally dense marker subsets, substantially reducing input dimensionality while retaining breed-discriminative genomic signal. Classification performance was subsequently evaluated across a portfolio of 13 ML algorithms — Logistic Regression, Ridge Regression, LASSO, Elastic Net, SVM-Linear, SVM-RBF, Random Forest, LightGBM, XGBoost, Gradient Boosting, PyTorch MLP, PyTorch CNN, and a Stacking Ensemble — alongside Kernel Ridge Regression (KRR) as a GBLUP-analogous parametric genomic baseline, spanning four model families (parametric, regularised, ensemble, and deep learning). This cross-species, multi-algorithm framework is intended to provide evidence-based method-selection guidance for breed traceability initiatives and conservation programmes operating under realistic data constraints.

2. Materials and Methods

2.1 Animal Material and Genomic Data

Cattle genotype data were obtained from the Dryad Digital Repository (doi:10.5061/dryad.th092), as originally published by Decker et al.[16]. Sheep genotype data were retrieved from the International Sheep Genomics Consortium (ISGC; https://www.sheephapmap.org/) [17]. Together, these publicly available resources yielded genotype data for 20 cattle breeds (n = 403) and 20 sheep breeds (n = 1,458), collectively representing a broad spectrum of genetic diversity. Cattle samples encompassed European taurine, indicine (Bos indicus), African sanga and tropical taurine, and Bali cattle (Bos javanicus) lineages, while sheep samples spanned European, African, Middle Eastern, American, and Central Asian populations. Breed composition and per-breed sample sizes are summarised in Table 2.

2.2 Genotype Quality Control, Imputation, and LD Pruning

Standard quality control (QC) was performed using PLINK 1.9 (Chang et al., 2015). SNP markers with a minor allele frequency (MAF) below 0.05 and individuals or markers with a

genotype call rate below 90% (missing rate >10%) were excluded. Following QC, missing genotypes were imputed using BEAGLE 5.0 (Browning et al., 2018), which performs simultaneous phasing and imputation via a hidden Markov model. After imputation, LD-based marker pruning was applied in PLINK 1.9 using a sliding-window approach (window size 50 SNPs, step 10 SNPs, r^2 threshold 0.70) to remove markers in strong LD, yielding a set of approximately independent SNPs for downstream analysis.

2.3 Autoencoder-Based Feature Selection

Given the high dimensionality of genome-wide SNP data, an autoencoder (AE) deep learning approach was employed prior to breed classification to identify the most informative marker subset; autoencoders have demonstrated utility for non-linear dimensionality reduction of SNP data, capturing population structure with greater fidelity than linear projections such as PCA [18,19]. The AE comprised an encoder–decoder architecture in which the encoder projected the input SNP vector through successive hidden layers of decreasing width into a low-dimensional bottleneck, and the decoder reconstructed the original input from this compressed representation [18]; batch normalisation, dropout regularisation, and early stopping (patience = 10 epochs; maximum 100 epochs) were applied throughout to mitigate overfitting. The pipeline supports binary, continuous, and categorical phenotype modes and was operated in categorical mode throughout this study, whereby breed identity is treated as a multi-class integer label enabling proper stratified k-fold assignment across all breed classes; the auxiliary supervised cross-entropy loss coefficient (α) was set to zero, preserving purely unsupervised reconstruction training. Stratified fold assignment in categorical mode ensures proportional breed representation in every training and validation partition, which is particularly important given the heterogeneous per-breed sample sizes in the present dataset.

Feature importance was quantified by combining two complementary scores: (i) a gradient-based (Jacobian) score computed as the mean absolute partial derivative of reconstruction loss with respect to each input SNP; (ii) an activation-based score calculated as the Pearson correlation between each input SNP and the corresponding bottleneck activation pattern. The two normalised scores were aggregated into a weighted composite and the top 5,000 SNPs were retained for all downstream classification analyses.

To prevent any information from held-out animals from influencing marker selection, autoencoder fitting and the top-5,000 selection were nested within the same cross-validation

used for classification (5-fold stratified, repeated over seeds 42, 43 and 44). Within each fold, the autoencoder was fitted only on the training partition; its hyperparameters were tuned by an inner cross-validation on that partition using Optuna (Tree-structured Parzen Estimation for the first 8 trials, then CMA-ES; median pruning, warm-up 3 trials; 25 trials). Feature-importance scores were computed from the fitted autoencoder, the top-5,000 SNPs were selected from the training partition alone, and this fold-specific marker set was applied to the corresponding held-out test partition. Held-out animals therefore contributed neither to autoencoder training, hyperparameter tuning, nor marker ranking, and the retained 5,000-SNP set is fold-specific rather than global.

2.4 Classification Algorithms

Breed discrimination was benchmarked across 14 classifiers.

Regularised linear models: Logistic Regression (Multinomial; L1/L2/Elastic Net penalties), Ridge Regression (L2), LASSO (L1), and Elastic Net (combined L1/L2).

Kernel-based classifiers: SVM-Linear and SVM-RBF; both with probability calibration via Platt scaling applied within the training fold.

Tree-based ensembles: Random Forest, Gradient Boosting, LightGBM, and XGBoost.

Deep learning classifiers: A PyTorch Multi-Layer Perceptron (MLP) with fully connected layers, batch normalisation, ReLU activations, dropout, and early stopping; and a PyTorch one-dimensional Convolutional Neural Network (CNN) with convolutional filters, batch normalisation, global average pooling, and a fully connected classification head.

Stacking Ensemble: Three base learners (Random Forest, Gradient Boosting, XGBoost) with out-of-fold predictions fed to a Multinomial Logistic Regression meta-learner trained via 5-fold internal cross-validation.

Kernel Ridge Regression (KRR), a GBLUP-analogous genomic baseline: Implemented via sklearn.kernel_ridge.KernelRidge, where the kernel function implicitly defines the genomic relationship structure and the ridge regularisation parameter (α) corresponds to the variance ratio. Class posteriors are obtained by softmax normalisation of one-vs-rest kernel scores and are not derived from a Bayesian probabilistic model; calibration results for this model are

specific to this implementation and should not be generalised to Bayesian GBLUP. Probability calibration was additionally applied to kernel ridge outputs to produce class posteriors.

2.5 Hyperparameter Optimisation

Model-specific hyperparameter spaces were explored using Bayesian sequential optimisation (BayesSearchCV, scikit-optimize) with 50 iterations per model. The objective was a composite score weighting overall accuracy (60%) and macro-averaged F1 (40%). Model-specific hyperparameter search spaces are summarised in Table 1; the selected (best) hyperparameter values are reported in Supplementary Table S1.

Table 1. Hyperparameter search spaces by model.

Model	Key Hyperparameters Optimised (n params)
Logistic Regression (Multinomial)	C, penalty {L1/L2/elasticnet}, solver, max_iter, tol, class_weight, l1_ratio (7)
Ridge Regression (Classification)	alpha, fit_intercept, solver, tol, max_iter, class_weight (6)
LASSO (Classification)	alpha, fit_intercept, max_iter, tol, warm_start, selection, class_weight (7)
Elastic Net (Classification)	alpha, l1_ratio, fit_intercept, max_iter, tol, warm_start, selection, class_weight (8)
SVM-Linear	C, class_weight, probability, tol, max_iter (5)
SVM-RBF	C, gamma, shrinking, class_weight, probability, tol, cache_size, max_iter (8)
Random Forest	n_estimators, max_depth, min_samples_split/leaf, max_features, min_impurity_decrease, max_samples, ccp_alpha, class_weight, bootstrap, criterion (11)
Gradient Boosting	n_estimators, learning_rate, max_depth, min_samples_split/leaf, subsample, max_features, min_impurity_decrease, ccp_alpha, validation_fraction, n_iter_no_change, tol (12)
LightGBM	n_estimators, learning_rate, max_depth, num_leaves, min_child_samples/weight, subsample, subsample_freq, colsample_bytree, min_gain_to_split, reg_alpha/lambda, boosting_type, max_bin (17)
XGBoost	n_estimators, learning_rate, max_depth, min_child_weight, subsample, colsample_bytree/bylevel/bynode, gamma, reg_alpha/lambda, max_delta_step, grow_policy (15)
PyTorch MLP	hidden_dims, dropout, lr, weight_decay, batch_size, epochs (6)
PyTorch CNN (1D)	n_filters, kernel_size, dropout, lr, weight_decay, batch_size, epochs (7)
Stacking Ensemble	meta-learner: Multinomial LR; base learners: RF, Gradient Boosting, XGBoost (fixed; internal 5-fold CV)

KRR	alpha, kernel {linear/rbf/polynomial}, gamma, degree, coef0 (via KernelRidge; 6)	

2.6 Cross-Validation Design and Data Leakage Prevention

Classification performance was estimated using 5-fold stratified cross-validation repeated across multiple random seeds (42, 43, 44), yielding 15 independent fold-level observations per model. Within each fold, 20% of animals formed the held-out test partition and 80% formed the training partition; every animal was tested exactly once per seed. All preprocessing operations — autoencoder-based marker selection, imputation, scaling, and oversampling — were encapsulated within scikit-learn Pipeline or ImbPipeline objects and fitted exclusively on the training partition, preventing leakage from test observations into any preprocessing step. As synthetic oversampling interpolates SNP dosages and may yield genotype values without a direct biological counterpart, it was applied only within the training partition of each fold, and held-out animals were never oversampled.

2.7 Class Imbalance Handling

Class imbalance across heterogeneous breed sizes was addressed at two levels. In the autoencoder pipeline, stratified fold assignment in categorical mode ensured each breed was proportionally represented in every training and validation partition. In the classification pipeline, a four-level mechanism was applied within training folds only: (1) automatic imbalance detection (ratio threshold 1.5); (2) synthetic oversampling in order of preference — ADASYN → Borderline-SMOTE → SMOTE — with adaptive nearest-neighbour selection and fallback to no oversampling when class sizes precluded synthetic generation; (3) class-weighted loss functions (class_weight='balanced'); and (4) scale_pos_weight adjustment for XGBoost and LightGBM.

2.8 Performance Metrics

Each model was evaluated on held-out test partitions using: overall accuracy, balanced accuracy (mean per-class recall), macro-averaged F1, precision, and recall; multi-class AUC-ROC (one-vs-rest, macro-averaged); Cohen's κ and MCC; Brier Score, ECE, MCE, Calibration Slope, and Log Loss. The train–test performance gap served as an overfitting diagnostic. All

estimates are fold-averaged; 95% bootstrap confidence intervals were computed for test accuracy.

2.9 Goodness-of-Fit (GoF) Analysis and Composite Model Ranking

Probability calibration was quantified through a panel of metrics computed on held-out test partitions: ECE, MCE, calibration slope (regression of empirical accuracy on predicted confidence), and log-loss. These metrics, together with AUC, Brier score, and cross-validation stability estimates, were compiled into a model performance heatmap for visual comparison across classifiers and species. Confusion matrices were generated for the two highest-ranked classifiers per species to identify breed-level misclassification patterns.

Overall model ranking was determined by two complementary approaches. First, a Composite Score was computed for each model as a weighted linear combination of seven normalised metrics: Accuracy (0.20), F1-Macro (0.20), AUC (0.15), Brier Score (0.15), Log Loss (0.15), ECE (0.10), and CV Accuracy (0.05). Each metric was normalised to [0, 1] before weighting, with lower-is-better metrics (Brier, Log Loss, ECE) inverted so that higher composite scores consistently indicate better overall performance.

Second, a Borda Count was computed as a rank-aggregation method independent of metric scale or normalisation assumptions. For each metric in the ranking set, models were assigned ordinal ranks (lower rank = better performance; for Calibration Slope, rank was assigned by absolute deviation from 1.0). The Borda score for each model was calculated as the sum of (n_models − rank) across all metrics, such that the model with the best aggregate rank position receives the highest Borda score. Confidence interval bounds were excluded from the Borda ranking to avoid double-counting. Together, the Composite Score and Borda Count provide convergent evidence for identifying the most performant classifiers across the multi-metric evaluation space.

2.10 Statistical Comparison of Models

Pairwise differences in classifier performance were assessed using the McNemar test applied to matched binary error vectors from held-out test partitions. For each model pair, McNemar's chi-squared statistic was computed from the counts of discordant predictions, and raw p-values were corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate

(BH-FDR) procedure across all 91 model pairs per species. A corrected p-value ≤ 0.05 was considered statistically significant; as predictions were aggregated across seeds, the counts of significant pairs are interpreted descriptively rather than as exact frequentist rates. Results are reported as pairwise significance matrices for each species. All tests were performed at $\alpha = 0.05$.

2.11 Software and Computational Environment

Both analysis pipelines were implemented in Python (version 3.9+). The autoencoder pipeline used PyTorch (≥2.0), Optuna (≥3.0), pandas-plink, scikit-learn, NumPy, and Pandas. The breed classification pipeline used scikit-learn (LogisticRegression, RidgeClassifier, Lasso, ElasticNet, SVC with linear and RBF kernels, RandomForestClassifier, GradientBoostingClassifier, StackingClassifier), scikit-optimize (BayesSearchCV), imbalanced-learn (SMOTE, ADASYN, BorderlineSMOTE), XGBoost, LightGBM, and PyTorch (MLP, CNN). The GoF analyzer used SciPy (statistical testing), statsmodels (multiple testing correction), and scikit-learn (calibration utilities). Visualisation was performed with Matplotlib and Seaborn. All random seeds were fixed for reproducibility. All computations were performed on a workstation equipped with an AMD Ryzen 9 processor and 32 GB of RAM.

3. Results

The dataset encompassed 40 breeds across two livestock species, with a total of 1,861 genotyped individuals. The cattle panel comprised 20 breeds (n = 403) representing three distinct species — *Bos taurus*, *Bos indicus*, and *Bos javanicus* — spanning origins from Western Europe, West Africa, South Asia, and Southeast Asia. The sheep panel was notably larger, with 20 breeds (n = 1,458), all classified as *Ovis aries*, drawn from a broad geographic range including the British Isles, continental Europe, the Middle East, East Africa, Central Asia, and Oceania. Whereas the cattle dataset maintained a uniform per-breed sample size for most breeds (n = 17–31), sheep breed representation was considerably more variable, ranging from under 40 to over 100 individuals per breed. This structural asymmetry between species — both in taxonomic diversity and sampling depth — provides a heterogeneous but realistic foundation for evaluating breed discrimination performance across contrasting population architectures (Table 2).

Table 2. Breed composition, species classification, geographic origin, and per-breed sample sizes for the cattle and sheep datasets used in this study.

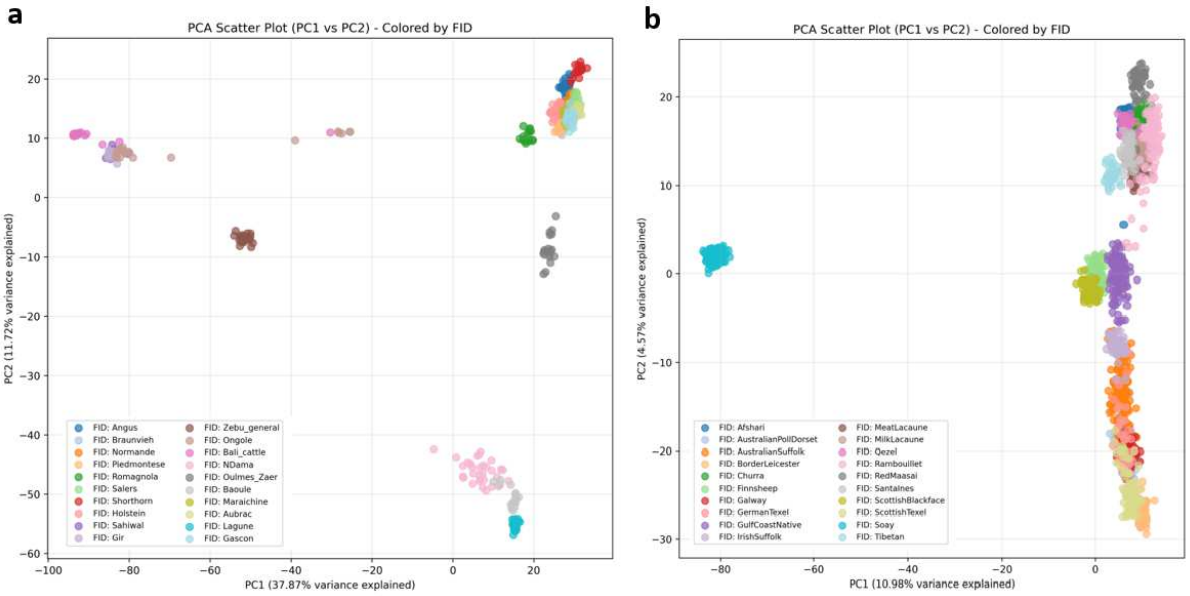
Cattle	*n*	Species	Native to	Sheep	*n*	Species	Native to
Angus	20	*Bos taurus*	Scotland, UK	Afshari	37	*Ovis aries*	Iran
Aubrac	20	*Bos taurus*	France	Australian Poll Dorset	108	*Ovis aries*	Australia
Bali cattle	20	*Bos javanicus*	Indonesia	Australian Suffolk	109	*Ovis aries*	Australia
Baoule	20	*Bos taurus*	West Africa	Border Leicester	48	*Ovis aries*	England/Scotland, UK
Braunvieh	20	*Bos taurus*	Switzerland	Churra	120	*Ovis aries*	Spain
Gascon	20	*Bos taurus*	France	Finnsheep	99	*Ovis aries*	Finland
Gir	20	*Bos indicus*	India	Galway	49	*Ovis aries*	Ireland
Holstein	20	*Bos taurus*	Cosmopolitan	German Texel	46	*Ovis aries*	Germany
Lagune	20	*Bos taurus*	West Africa	Gulf Coast Native	94	*Ovis aries*	USA
Maraichine	19	*Bos taurus*	France	Irish Suffolk	55	*Ovis aries*	Ireland
N'Dama	31	*Bos taurus*	West Africa	Meat Lacaune	78	*Ovis aries*	France
Normande	20	*Bos taurus*	France	Milk Lacaune	103	*Ovis aries*	France
Ongole	20	*Bos indicus*	India	Qezel	35	*Ovis aries*	Iran / Azerbaijan
Oulmes Zaer	19	*Bos taurus*	Morocco	Rambouillet	102	*Ovis aries*	France
Piedmontese	20	*Bos taurus*	Italy	Red Maasai	45	*Ovis aries*	East Africa
Romagnola	20	*Bos taurus*	Italy	SantaInes	47	*Ovis aries*	Brazil
Sahiwal	17	*Bos indicus*	South Asia	Scottish Blackface	56	*Ovis aries*	Scotland, UK
Salers	20	*Bos taurus*	France	Scottish Texel	80	*Ovis aries*	Scotland, UK
Shorthorn	17	*Bos taurus*	England, UK	Soay	110	*Ovis aries*	Scotland, UK
Zebu	20	*Bos indicus*	South Asia	Tibetan	37	*Ovis aries*	Tibet, China
Total	**403**				**1458**		

The PCA revealed markedly contrasting population architectures between the two species. In cattle (Figure 1a), the first two principal components jointly captured nearly half of the total genotypic variance. The most pronounced axis of differentiation (PC1) broadly separated *Bos indicus* and *Bos javanicus* lineages from *Bos taurus* breeds, while within-*Bos taurus* variation was further partitioned along PC2, distinguishing European from African breeds. Bali cattle (*Bos javanicus*) occupied a distinctly peripheral position, consistent with its divergent taxonomic status.

In contrast, the sheep panel (Figure 1b) exhibited substantially compressed cluster separation, with the first two PCs collectively explaining less than one-sixth of total variance — a roughly three-fold reduction relative to cattle. This compression is consistent with greater allelic continuity across *Ovis aries* breeds and suggests a more complex, multi-dimensional population structure that is not captured by the leading components alone; differences in array ascertainment between the two species may also contribute (see Discussion). Several geographically or phenotypically divergent breeds — including Soay and Tibetan — showed comparatively greater separation from the main cluster mass, whereas commercially selected and regionally proximate breeds exhibited considerable overlap. This fundamental difference

360 in PCA resolution between species established the analytical context for subsequent machine
361 learning–based breed discrimination.

362 **Figure 1.** Principal component analysis (PCA) of 50K genotypic variation in cattle (a) and
363 sheep (b) datasets, with individuals colored by breed.



364

Across both species, the top-performing classifiers achieved high discrimination accuracy, yet the two datasets revealed markedly different patterns of inter-model variation (Table 3). In cattle, linear and regularised regression approaches — including Logistic Regression, Ridge, LASSO, Elastic Net, and KRR — clustered at the top of the performance hierarchy alongside PyTorch MLP, with the leading group achieving near-ceiling discrimination across all reported metrics. Performance then declined progressively through kernel-based and tree-ensemble methods, with Gradient Boosting recording the lowest accuracy — a gap of more than 25 percentage points separating the top and bottom of the cattle leaderboard. This pronounced stratification indicates that the high-dimensional linear structure of SNP data strongly favours linear decision boundaries in cattle breed separation.

In sheep, the overall performance ceiling was marginally lower, yet the inter-model spread was substantially compressed: all 14 models exceeded 0.96 accuracy, and the range from best to worst was less than 3 percentage points. Notably, tree-based boosting methods — which ranked at the bottom in cattle — performed considerably more competitively in sheep, approaching the accuracy of linear models.

Across both species, MLP maintained consistently strong performance, whereas the PyTorch CNN lagged behind MLP in cattle despite comparable outcomes in sheep, pointing to an architecture-dataset interaction effect. KRR achieved accuracy on par with the best ML models in both species; however, its substantially elevated Brier scores relative to near-zero values for probabilistic classifiers such as logistic regression and PyTorch MLP indicate poor probabilistic output quality despite high discriminative accuracy — a distinction with practical implications for confidence-based assignment. Specificity exceeded 0.990 universally; under one-vs-rest evaluation of 20 classes where true negatives dominate, near-unity specificity is expected and offers limited discriminative information here.

Table 3. Classification performance of 14 machine learning and statistical models evaluated on held-out test sets for cattle (upper) and sheep (lower) breed discrimination tasks, ranked by accuracy within each species.

Cattle Test Metrics

Model	Accuracy	F1 Macro	Precision	Recall	Specificity	Kappa	MCC	Balanced Accuracy	Brier
Logistic Regression	0.9975	0.9975	0.9976	0.9975	0.9999	0.9974	0.9974	0.9975	0.0004
Ridge Regression	0.9975	0.9975	0.9976	0.9975	0.9999	0.9974	0.9974	0.9975	0.0342
PyTorch MLP	0.9975	0.9975	0.9976	0.9975	0.9999	0.9974	0.9974	0.9975	0.0002
KRR	0.9975	0.9975	0.9976	0.9975	0.9999	0.9974	0.9974	0.9975	0.0425
LASSO	0.9950	0.9954	0.9961	0.9950	0.9997	0.9948	0.9948	0.9950	0.0038
SVM Linear	0.9950	0.9950	0.9952	0.9950	0.9997	0.9948	0.9948	0.9950	0.0101
Elastic Net (Classification)	0.9950	0.9954	0.9961	0.9950	0.9997	0.9948	0.9948	0.9950	0.0026
SVM RBF	0.9876	0.9879	0.9893	0.9874	0.9993	0.9869	0.9870	0.9874	0.0103
Random Forest	0.9603	0.9598	0.9636	0.9596	0.9979	0.9582	0.9584	0.9596	0.0199
PyTorch CNN	0.9156	0.9148	0.9270	0.9143	0.9956	0.9111	0.9117	0.9143	0.0065
LightGBM	0.9107	0.9123	0.9164	0.9124	0.9953	0.9059	0.9061	0.9124	0.0062
Stacking Ensemble	0.8660	0.8684	0.8728	0.8676	0.9929	0.8588	0.8590	0.8676	0.0119
XGBoost	0.8437	0.8439	0.8482	0.8453	0.9918	0.8353	0.8356	0.8453	0.0121
Gradient Boosting	0.7395	0.7378	0.7518	0.7414	0.9863	0.7256	0.7264	0.7414	0.0207

Sheep Test Metrics

Model	Accuracy	F1 Macro	Precision	Recall	Specificity	Kappa	MCC	Balanced Accuracy	Brier
Elastic Net	0.9925	0.9923	0.9924	0.9924	0.9996	0.9920	0.9920	0.9924	0.0009
Logistic Regression	0.9918	0.9918	0.9919	0.9919	0.9996	0.9913	0.9913	0.9919	0.0007
Ridge Regression	0.9918	0.9916	0.9918	0.9915	0.9996	0.9913	0.9913	0.9915	0.0304
KRR	0.9918	0.9916	0.9918	0.9915	0.9996	0.9913	0.9913	0.9915	0.0414
PyTorch MLP	0.9911	0.9915	0.9913	0.9919	0.9995	0.9905	0.9906	0.9919	0.0008
SVM Linear	0.9904	0.9904	0.9909	0.9903	0.9995	0.9898	0.9898	0.9903	0.0101
SVM RBF	0.9904	0.9904	0.9913	0.9897	0.9995	0.9898	0.9898	0.9897	0.0101
LASSO	0.9904	0.9902	0.9901	0.9906	0.9995	0.9898	0.9898	0.9906	0.0010
Stacking Ensemble	0.9822	0.9813	0.9829	0.9798	0.9991	0.9811	0.9811	0.9798	0.0017
PyTorch CNN	0.9822	0.9800	0.9799	0.9806	0.9991	0.9811	0.9811	0.9806	0.0014
LightGBM	0.9760	0.9749	0.9759	0.9743	0.9987	0.9745	0.9745	0.9743	0.0018
Gradient Boosting	0.9746	0.9734	0.9758	0.9715	0.9986	0.9731	0.9731	0.9715	0.0020
XGBoost	0.9712	0.9681	0.9697	0.9667	0.9985	0.9694	0.9694	0.9667	0.0030
Random Forest	0.9684	0.9677	0.9710	0.9656	0.9983	0.9665	0.9666	0.9656	0.0177

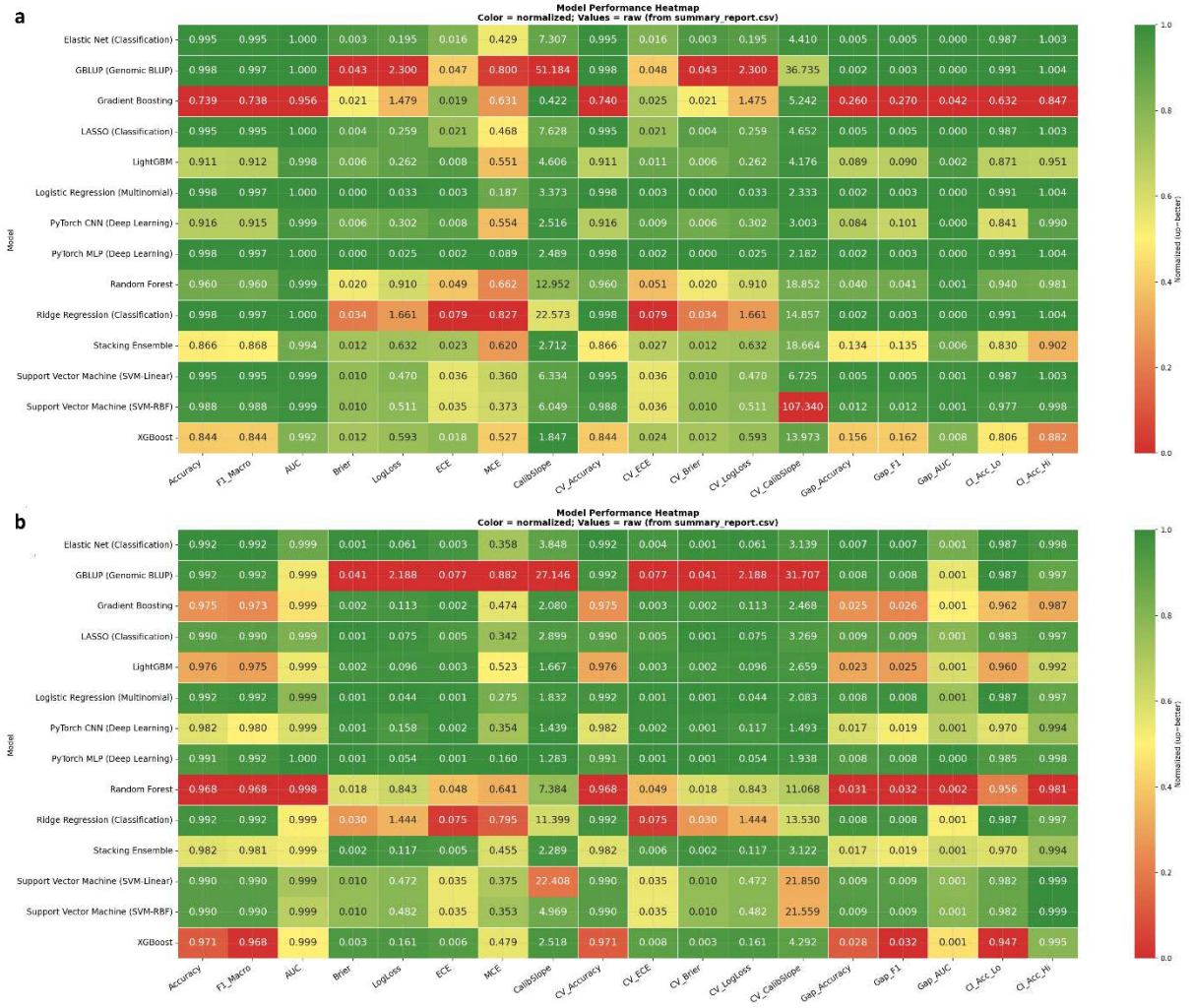
Figure 2 extends the evaluation framework to discrimination capacity (AUC), probability calibration (ECE, MCE, calibration slope, log-loss), cross-validation stability, and generalisation gap — metrics absent from Table 3 that jointly determine the operational reliability of each classifier.

AUC revealed a species-dependent stratification pattern. In cattle (Figure 2a), logistic regression, LASSO, elastic net, KRR, ridge regression, and PyTorch MLP all achieved AUC = 1.000, forming a discrete ceiling tier, while gradient boosting descended sharply to 0.956 — a gap of more than four percentage points separating boosting-family models from their linear counterparts. In sheep (Figure 2b), AUC compression was pronounced: all 14 classifiers remained within the 0.998–1.000 interval, with PyTorch MLP at the apex, confirming that the sheep population's genomic structure is more uniformly separable across algorithmic families.

Probability calibration exposed a consistent dissociation between discriminative and probabilistic performance conserved across both species. KRR produced calibration slopes of 51.184 and 27.146 in cattle and sheep respectively, alongside log-loss values of 2.300 and 2.188, indicating poor calibration. Because empirical accuracy and predicted confidence are both bounded in [0,1], slope values of this magnitude reflect near-constant predicted-confidence distributions rather than a graded severity of miscalibration, and should be read as flags rather than on a linear scale. Ridge regression showed the same signature (cattle CalibSlope: 22.573, log-loss: 1.661), confirming that penalty-based shrinkage distorts posterior probability estimates even when class boundary placement is optimal. In contrast, PyTorch MLP returned ECE values of 0.002 and 0.001 and log-loss values of 0.025 and 0.054 for cattle and sheep respectively, achieving the best-calibrated posteriors among all evaluated models. Logistic regression closely followed (cattle ECE: 0.003; sheep ECE: 0.001), establishing that probabilistic classifiers with explicit likelihood optimisation provide reliable confidence estimates where regularised regressors do not.

Generalisation gap analysis revealed a striking species asymmetry. In cattle, gradient boosting exhibited train-to-test accuracy and F1 gaps of 0.260 and 0.270 respectively, and XGBoost gaps of 0.156 and 0.162 — patterns indicative of severe overfitting to training data in the high-dimensional cattle SNP space. Stacking ensemble accuracy (gap: 0.134) and LightGBM (gap: 0.089) occupied an intermediate zone. By contrast, all linear and probabilistic models maintained gaps ≤ 0.005 in cattle, demonstrating that their training and held-out estimates are interchangeable. In sheep, the gap landscape collapsed uniformly: even Random Forest, Gradient Boosting, and XGBoost gaps were confined to 0.025–0.031, suggesting that the more homogeneous population structure in sheep substantially dampens the overfitting tendency of tree-based ensembles.

Figure 2. Detailed Model Performance Heatmaps — Cattle (a) and Sheep (b)

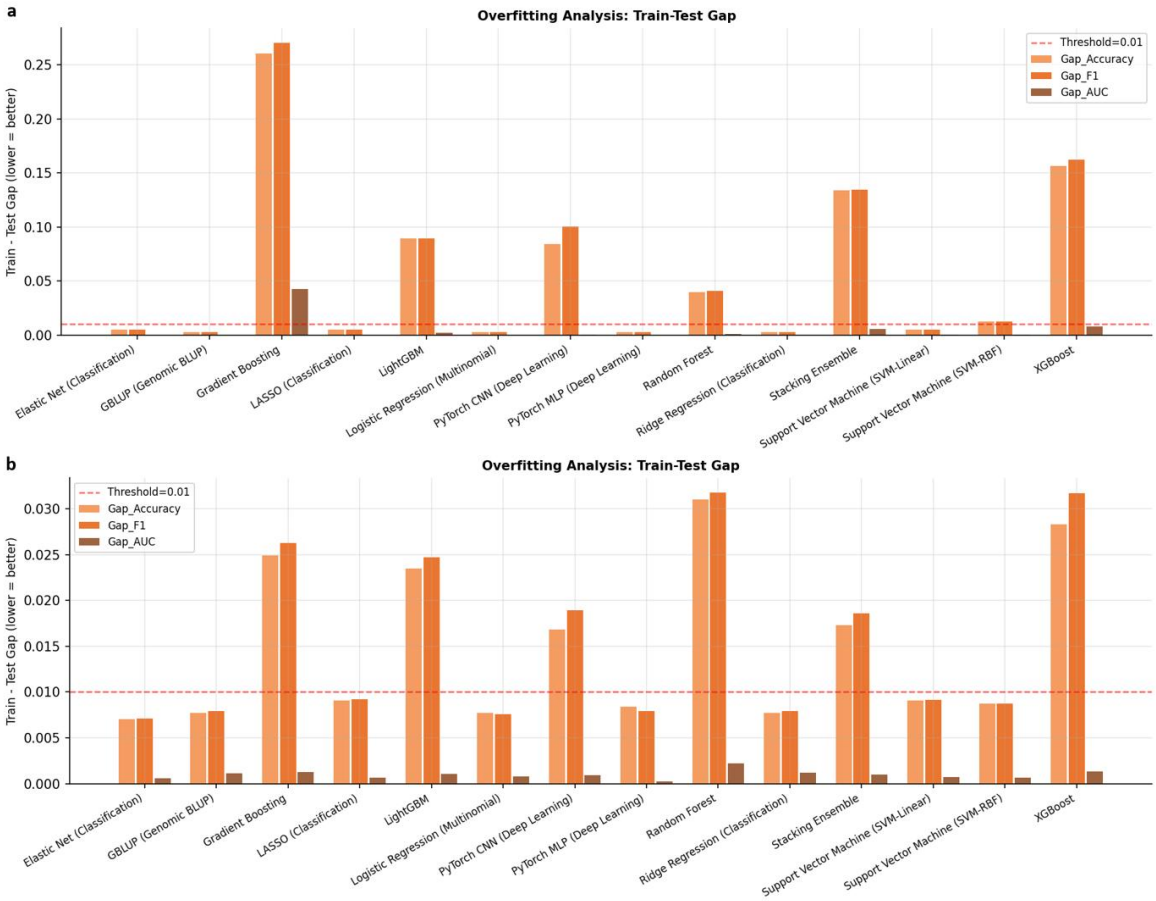


Accuracy: proportion of correctly classified individuals across all breed classes. *F1 Macro*: unweighted mean of per-class F1 scores, equally penalising precision and recall failures regardless of class size. *Precision*: fraction of predicted breed assignments that are correct. *Recall*: fraction of true breed members correctly identified; equivalent to sensitivity. *Specificity*: proportion of non-members correctly excluded per class. *Kappa*: Cohen's kappa coefficient, quantifying agreement between predicted and true labels after correcting for chance. *MCC*: Matthews Correlation Coefficient; a single balanced measure robust to class-size asymmetry, ranging from −1 (inverse prediction) to +1 (perfect prediction). *Balanced Accuracy*: arithmetic mean of per-class recall, insensitive to breed-size imbalance. *Brier*: mean squared error between predicted class probabilities and true binary outcomes; lower values indicate sharper, better-calibrated posteriors. *AUC*: area under the receiver operating characteristic curve, summarising discrimination capacity across all classification thresholds. *LogLoss*: negative log-likelihood of predicted probabilities; penalises confident mis-assignments exponentially. *ECE*: Expected Calibration Error; average absolute deviation between predicted confidence and empirical accuracy across probability bins. *MCE*: Maximum Calibration Error; worst-case bin-level miscalibration. *CalibSlope*: slope of the regression of empirical accuracy on predicted confidence. *CV_Accuracy / CV_ECE / CV_Brier / CV_LogLoss / CV_CalibSlope*: five-fold cross-validation estimates of the corresponding test metrics, used to assess stability of model performance across data partitions. *Gap_Accuracy* / Gap_F1 / Gap_AUC: absolute difference between training and held-out test performance for each metric; values approaching zero indicate negligible overfitting. *CI_Acc_Lo / CI_Acc_Hi*: lower and upper bounds of the 95% bootstrap confidence interval for test accuracy.

The bar charts in Figure 3 contextualise the generalisation gap metrics through direct visual comparison against a 0.01 threshold, above which overfitting is considered operationally meaningful. The most informative feature of Figure 3 is not the individual model values —

reported in full in Figure 2 — but the contrast in y-axis scale between panels: the cattle panel (Figure 3a) spans an order of magnitude wider than its sheep counterpart (Figure 3b), capturing in a single visual the species asymmetry that defines the overfitting landscape of this study. In cattle, boosting-family models breach the threshold conspicuously, whereas the majority of classifiers remain sub-threshold; in sheep, the threshold itself becomes a near-universal ceiling, with virtually all models compressed beneath it regardless of algorithmic family. Gap_AUC remains negligible across both species and all models, indicating that rank-ordering discrimination is preserved under training conditions even where absolute accuracy degrades at test time.

Figure 3. Overfitting Analysis: Train-Test Gap — Cattle (a) and Sheep (b)



Kernel ridge regression (KRR) was implemented using a kernel-based linear approach with a one-vs-rest multiclass decomposition. The kernel matrix was computed exclusively from the training partition within each cross-validation fold, ensuring that no information from held-out samples contributed to model fitting — a design consistent with the leakage-free preprocessing applied to all other classifiers in the pipeline. Class probabilities were obtained by applying

softmax normalisation to the raw kernel scores generated under the one-vs-rest decomposition. While this procedure yields mathematically valid probability distributions, they are not derived from an explicit probabilistic model. Accordingly, calibration-sensitive metrics such as ECE, Brier score, and calibration slope should be interpreted with appropriate caution for this model; discrimination metrics including accuracy, macro F1-score, and AUC retain full interpretational validity.

Per-class recall analysis revealed pronounced breed-level heterogeneity in cattle and near-uniform discrimination in sheep, a contrast that macro-averaged metrics in Table 3 partially obscure.

In cattle (Figure 4a), no breed achieved perfect recall universally across all 14 classifiers, with sub-threshold recall values confined exclusively to boosting-family models. Aubrac cattle emerged as the most challenging breed, with recall collapsing to 0.45 under gradient boosting while reaching 1.000 under linear classifiers — a within-breed range of 0.55 across models. Romagnola, Piedmontese, and Salers presented comparable model-sensitivity, with minimum recalls of 0.50, 0.55, and 0.55 respectively. Braunvieh was the sole breed for which no classifier achieved perfect recall, with a ceiling of 0.95 across all 14 models. As Braunvieh comprised 20 animals, this ceiling corresponds to a single consistently misassigned individual rather than a breed-wide deficit; it is consistent with, though does not on a single-animal basis establish, genomic proximity to neighbouring European taurine breeds. Gradient boosting returned sub-0.90 recall for 15 of 20 breeds, while all linear and probabilistic classifiers maintained breed-level recall ≥ 0.95 throughout.

In sheep (Figure 4b), the recall floor held at 0.846 across all 280 breed–model combinations, with not a single cell falling below 0.80. Four breeds — Finnsheep, Scottish Blackface, Soay, and Tibetan — achieved perfect recall of 1.000 across every classifier, and 98.6% of all cells remained at or above 0.90. The most challenging sheep breed, Meat Lacaune, dipped below 0.90 in only one of 14 models, a contrast to the systematic recall failures observed for multiple cattle breeds. Even the worst-performing sheep classifier (Random Forest, minimum recall 0.846) outperformed the worst-performing cattle model (Gradient Boosting, minimum 0.450) by a margin exceeding 0.39.

Figure 4. Per-Class Recall Heatmaps — Cattle (a) and Sheep (b)

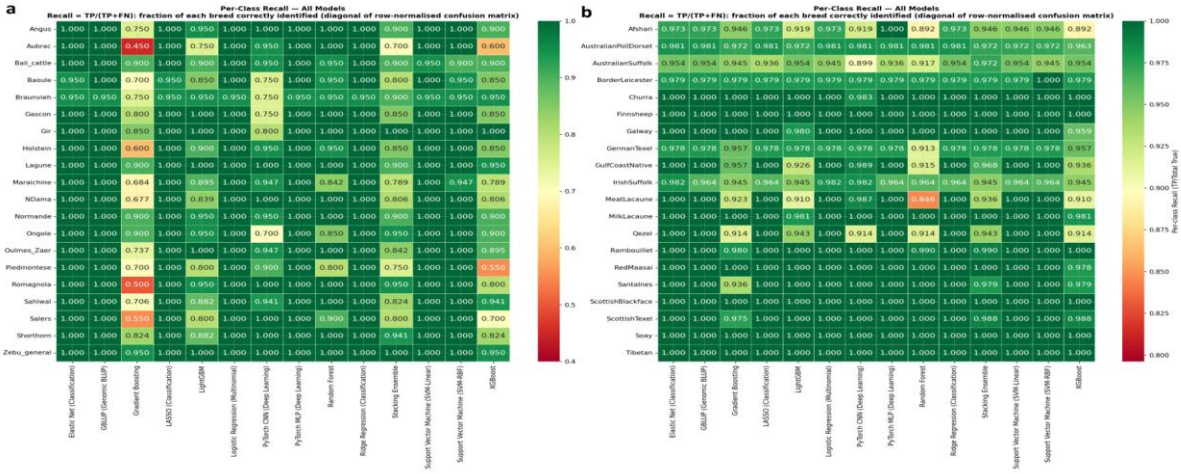


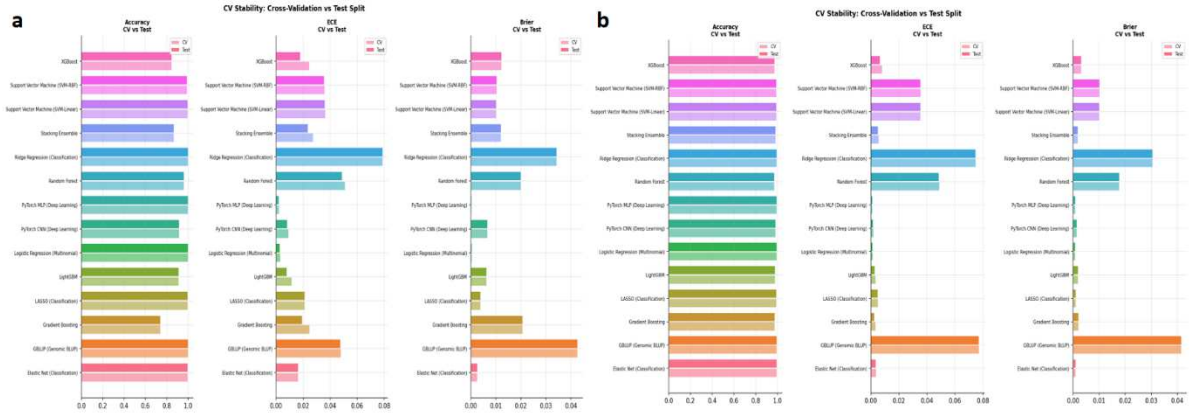
Figure 5 presents CV and test measurements of accuracy, ECE, and Brier as paired horizontal bars, with near-perfect visual overlap serving as the primary indicator of cross-validation reliability.

In the accuracy panel, CV and test bars are visually indistinguishable across all classifiers in both species — no model produces a perceptible gap between the two estimates, confirming that accuracy-based model ranking derived from cross-validation faithfully reflects held-out performance.

The ECE panel reveals a more informative pattern. For the large majority of classifiers, bar lengths remain closely matched between CV and test. Ridge Regression and KRR, however, stand out not for divergence between their CV and test bars — which remain tightly paired — but for the conspicuous length of both bars relative to all other models. The visual alignment of these elongated pairs confirms that their elevated miscalibration is consistently reproduced across data partitions, identifying it as a structural rather than stochastic property. A modest visual separation between CV and test ECE bars is discernible for select boosting-family classifiers in cattle, an effect that largely disappears in the sheep panel.

In the Brier panel, Ridge Regression and KRR produce the longest bars in cattle and sheep, with CV and test extending to a comparable degree. Across all remaining models in both species, Brier bars are visually negligible and mutually overlapping, indicating that probability sharpness estimates are stable across partitions regardless of model class.

Figure 5. CV Stability: Cross-Validation vs Test Split — Cattle (a) and Sheep (b)



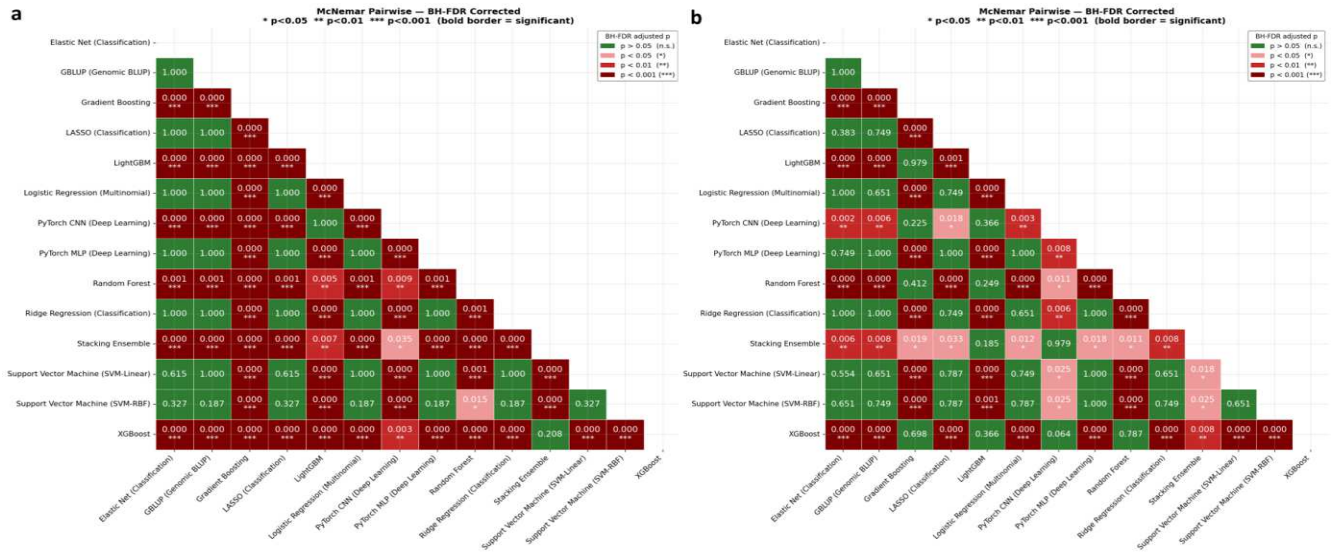
Pairwise McNemar tests with Benjamini–Hochberg FDR correction formalised the statistical structure of performance differences across 91 model pairs per species.

In cattle (Figure 6a), 61 of 91 pairs (67.0%) reached significance, revealing a sharply bipartite classifier landscape. A top tier whose within-group pairwise differences were not statistically significant comprised elastic net, KRR, LASSO, logistic regression, PyTorch MLP, ridge regression, and both SVM variants — all within-group comparisons returned p_adj = 1.000, indicating that their prediction errors are statistically indistinguishable at the individual-sample level. Each of these models reached significance against exactly 6 of 13 comparators, exclusively those in the lower performance tier. In contrast, gradient boosting and random forest were each significantly different from all 13 remaining classifiers, occupying statistically isolated positions — gradient boosting at the performance floor and random forest as a mid-tier outlier distinguishable from both the linear top tier and the lower boosting cluster. LightGBM, PyTorch CNN, stacking ensemble, and XGBoost occupied an intermediate zone with 12 significant pairs each; notably, stacking and XGBoost were mutually non-significant (p_adj = 0.208), and LightGBM and PyTorch CNN were indistinguishable from each other (p_adj = 1.000).

In sheep (Figure 6b), 52 of 91 pairs (57.1%) were significant — a reduction consistent with the overall compression of inter-model performance differences in that species. The linear and probabilistic top tier was preserved intact, with all within-group comparisons remaining non-significant. The most consequential species-level contrast was observed in the boosting family:

gradient boosting, which was significantly different from all 13 peers in cattle, became statistically indistinguishable from LightGBM, PyTorch CNN, random forest, and XGBoost in sheep — its significant pair count falling from 13 to 9. Random Forest remained highly statistically isolated in sheep (10/13 significant pairs), confirming its structural distinctiveness even where its absolute performance approached the boosting cluster. Stacking ensemble registered the highest significant pair count in sheep (11/13), driven by its divergence from both the linear top tier and the boosting cluster, reflecting its intermediate but statistically differentiated position.

Figure 6. McNemar Pairwise Tests — Cattle (a) and Sheep (b)



Model rankings were derived from two complementary frameworks: a weighted composite score integrating normalised discrimination and calibration metrics (accuracy and F1: 0.20 each; AUC, Brier, log-loss: 0.15 each; ECE: 0.10; CV accuracy: 0.05), and a weight-agnostic Borda count assigning equal influence to all metrics. The composite score constitutes the primary ranking criterion; Borda serves as a corroborative reference (Table 4).

PyTorch MLP ranked first under both frameworks in cattle and retained the top composite position in sheep, representing the sole model achieving consistent cross-species primacy. In cattle, logistic regression and elastic net occupied ranks 2–3 by composite score, with LASSO immediately following — a contiguous linear tier whose internal Borda ordering closely mirrored composite ranking. Gradient Boosting placed last in both frameworks in cattle (composite: 0.21), the widest margin from the penultimate model observed in either species.

The most informative species contrast appeared in the mid-table: ridge regression and KRR, which ranked 10th and 12th by composite in cattle, fell further to 9th and 12th in sheep despite improved absolute performance — their persistently high Brier and ECE values systematically suppressing their weighted scores. Random Forest presented the sharpest Borda–composite divergence in cattle, ranking 9th by composite but 13th by Borda, reflecting that its strong accuracy is offset by poor calibration-metric ranks when all metrics carry equal weight. In sheep, gradient boosting recovered to 11th (composite: 0.57) from 14th in cattle (0.21), the largest rank improvement across species, consistent with the reduced overfitting observed in that population.

Table 4. Composite and Borda Model Rankings — Cattle and Sheep

Cattle					Sheep				
Model	**Comp. score**	**Comp. Rank**	**Borda score**	**Borda Rank**	**Model**	**Comp. score**	**Comp. Rank**	**Borda score**	**Borda Rank**
PyTorch MLP	1.00	1	12.16	1	PyTorch MLP	0.98	1	11.44	2
Logistic Regression	1.00	2	11.47	2	Elastic Net	0.98	2	10.56	3
Elastic Net	0.96	3	9.25	3	Logistic Regression	0.96	3	11.50	1
LASSO	0.94	4	8.19	4	LASSO	0.93	4	8.25	4
SVM-Linear	0.89	5	6.56	6	SVM-Linear	0.81	5	5.31	9
SVM-RBF	0.87	6	5.56	8	SVM-RBF	0.80	6	5.63	8
PyTorch CNN	0.80	7	6.81	5	Stacking Ensemble	0.72	7	6.03	7
LightGBM	0.80	8	6.38	7	PyTorch CNN	0.72	8	7.72	5
Random Forest	0.74	9	3.38	13	Ridge Regression	0.61	9	4.50	11
Ridge Regression	0.67	10	5.53	9	LightGBM	0.61	10	6.25	6
Stacking Ensemble	0.64	11	3.63	12	Gradient Boosting	0.57	11	5.25	10
KRR	0.64	12	5.34	10	KRR	0.52	12	3.88	12
XGBoost	0.61	13	4.00	11	XGBoost	0.48	13	3.38	13
Gradient Boosting	0.21	14	2.75	14	Random Forest	0.22	14	1.31	14

Comp. Score: Composite score; Borda Score: normalised mean of per-metric reverse ranks across all models.

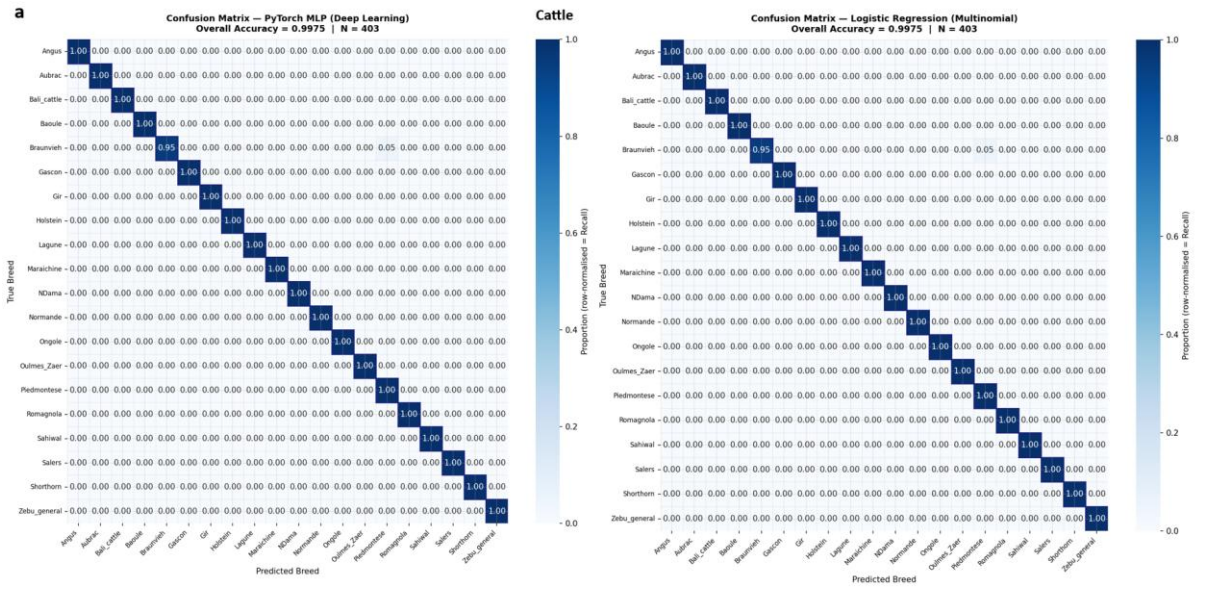
Confusion matrices for the two highest-ranked classifiers per species provide breed-level resolution of classification errors that aggregate metrics cannot capture.

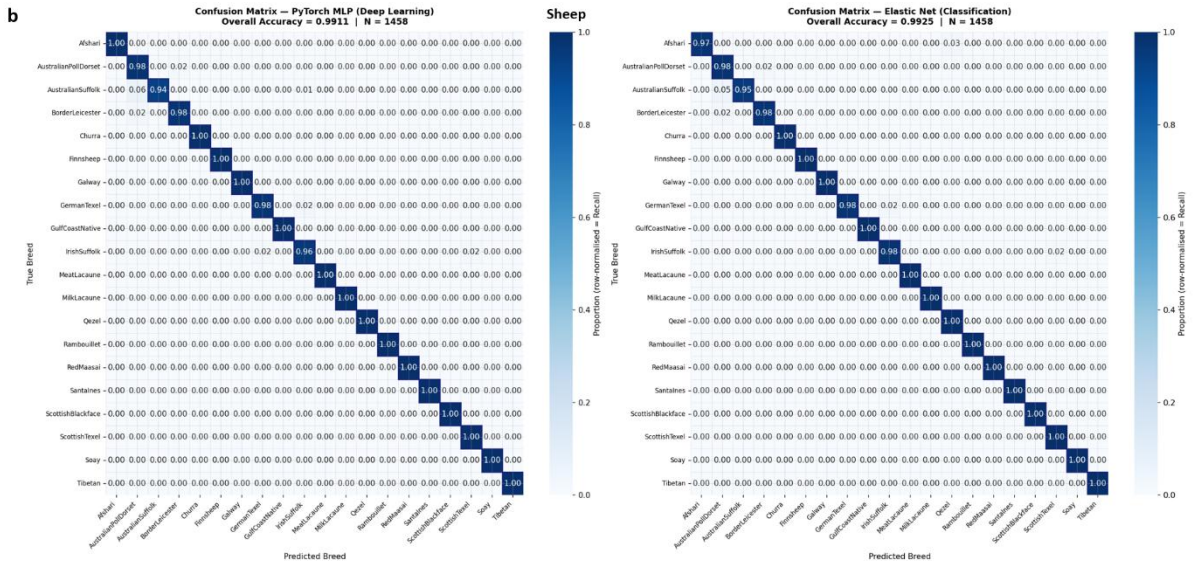
In cattle (Figure 7a), both PyTorch MLP and logistic regression correctly assigned 19 of 20 breeds with perfect recall, with Braunvieh constituting the sole shared error source for both models (0.95 recall). The complete identity of error location between two algorithmically distinct classifiers is notable: it indicates that the residual misclassification is population-driven rather than model-specific, consistent with the PCA landscape in which within-*Bos taurus*

variation along PC2 places Braunvieh in close proximity to neighbouring European taurine breeds, leaving a genomic margin too narrow for perfect separation even at high overall accuracy. No other breed produced off-diagonal mass in either matrix.

In sheep (Figure 7b), five breeds were imperfectly recalled by both PyTorch MLP and elastic net: Australian Suffolk, Australian Poll Dorset, Border Leicester, German Texel, and Irish Suffolk. This shared error profile points to a structurally determined confound rather than algorithmic limitation. All five are commercially selected European-derived breeds with documented genetic proximity, and their sub-perfect recall aligns directly with the PCA finding that regionally proximate and commercially selected sheep breeds exhibit considerable cluster overlap under the leading components — a reflection of the compressed, multi-dimensional population structure that characterises the sheep dataset. Within this common error set, Australian Suffolk produced the lowest recall in both models (MLP: 0.94; elastic net: 0.95), while Irish Suffolk showed the largest between-model divergence: MLP recalled it at 0.96 versus elastic net's 0.98, suggesting a marginal advantage of the elastic net in resolving breeds with overlapping allele frequency profiles. Conversely, Afshari was misclassified exclusively by elastic net (0.97) and recalled perfectly by MLP, representing the sole breed where the models diverged in error direction. Breeds identified by PCA as structurally distinct — Soay, Tibetan, Finnsheep, and Scottish Blackface — achieved perfect recall under both classifiers, confirming that PCA-resolved genomic distance is strongly associated with classification accuracy.

Figure 7. Confusion Matrices for Top-Ranked Models — Cattle (a) and Sheep (b)





4. Discussion

The classification pipeline employed in this study incorporated an autoencoder-based feature extraction stage to reduce the raw genome-wide SNP input space to a compact representation of the most informative markers prior to model training. This compression preserved breed-discriminative genomic signal while substantially reducing input dimensionality, an approach that aligns with a growing body of evidence supporting deep learning architectures as effective tools for non-linear dimensionality reduction of genotype data. Ausmees and Nettelblad[18] demonstrated that convolutional autoencoders applied to SNP data are capable of capturing population structure with greater fidelity than standard linear projection methods such as PCA, while retaining the spatial properties of linkage disequilibrium across the genome. Complementarily, López-Cortés et al.[19] showed that deep autoencoder-based preprocessing of SNP data improved individual assignment accuracy relative to PCA when combined with clustering algorithms, suggesting that non-linear compression of genomic feature spaces may yield representations more amenable to downstream classification. In the present study, the retention of near-ceiling discriminative performance across multiple model families following this dimensionality reduction step indicates that the autoencoder-derived feature space retained sufficient inter-breed structure, though the extent to which compression itself contributed to — rather than simply preceded — this outcome warrants further investigation with controlled ablation designs.

4.1 Population Structure as the Primary Factor Associated with Classification Performance

The most consistent finding across the present results is that the degree of genomic differentiation among breeds, as reflected in PCA resolution, corresponded closely to inter-model agreement and overall classification difficulty. In cattle, the taxonomic heterogeneity of the dataset — spanning three Bos species — produced well-separated population clusters in low-dimensional space, and this structural clarity translated into near-uniform ceiling performance among linear and probabilistic classifiers. In sheep, the allelic continuity characteristic of Ovis aries compressed PCA resolution to less than one-sixth of total variance in the first two components, yet classification accuracy remained high across most models, suggesting that breed-discriminative information is encoded in higher-order dimensions of the SNP space even when standard PCA fails to surface it. This pattern is broadly consistent with the observation by Farooq et al.[20] that population structure exerts a meaningful influence on the relative performance of machine learning approaches in genomic data, with structured populations tending to favour linear models. The present results extend this principle to a multi-species breed discrimination context, where inter-species taxonomic distance may be associated with the advantage of linear decision boundaries. An important caveat applies to these cross-species comparisons: the cattle and sheep datasets were generated on different commercial SNP arrays (BovineSNP50 and OvineSNP50), whose marker ascertainment schemes differ. SNP ascertainment bias is known to influence apparent population structure and the variance captured by leading principal components, so the contrast in PCA resolution between species cannot be attributed to population architecture alone. Zhao et al.[2] similarly noted that breed identification accuracy in cattle was robust across a range of classification methods, pointing to the underlying genetic differentiation among breeds as the primary enabler of high accuracy rather than any specific algorithmic choice.

4.2 The Primacy of Linear Models in High-Dimensional Genomic Classification

The dominance of linear and regularised classifiers across both species is broadly interpretable through the additive structure of SNP data. Returning to the question raised in the Introduction — whether breed discrimination depends on non-linear or additive signal — the present results indicate that, in these datasets, breed boundaries were largely captured by additive allele-frequency differences. When breed boundaries are largely determined by allele frequency differences that are linearly distributed across the genome, linear decision boundaries are

sufficient and parsimonious. Farooq et al.[20] reported that ensemble machine learning methods offered advantages primarily when a substantial proportion of phenotypic variance was attributable to non-linear allele effects, whereas for largely linear genetic architectures their performance was comparable to linear parametric models.

Notably, Thorsrud et al.[21] reported that GBLUP performed comparably to Random Forest, SVM, XGBoost, and MLP in genomic prediction of binary traits in working dogs, concluding that simpler models requiring less parameter optimisation may be sufficient in many genomic classification contexts. The present findings are consistent with this view, particularly in sheep where inter-model performance differences were small throughout.

4.3 Calibration–Accuracy Dissociation and Its Implications for Applied Breed Assignment

One of the more consequential findings in this study is the systematic dissociation between discriminative accuracy and probability calibration for specific model classes. Models that achieve near-ceiling classification accuracy do not necessarily produce reliable posterior probability estimates, and this distinction carries practical significance for applications in which breed assignment decisions are conditioned on confidence thresholds — such as pedigree verification, import-export certification, or genomic traceability systems. Dormann[22] noted that raw probability predictions from machine learning classifiers may deviate substantially from true empirical frequencies, particularly for models that optimise non-likelihood target functions, and recommended calibration regression as a corrective step. This concern applies directly to ridge regression in this study, where L2 shrinkage distorts the correspondence between predicted scores and empirical class frequencies. For KRR, miscalibration arises from a distinct source: class posteriors are obtained by softmax normalisation of kernel scores that are not derived from a probabilistic generative model, making direct application of calibration diagnostics designed for probabilistic ML classifiers methodologically imprecise.

For applications where calibrated probability output is required, post-hoc recalibration methods represent a practical remedy. Berisha et al.[23] demonstrated that logistic and beta calibration approaches consistently improved probability estimates across a range of classifiers, with regression on logit-transformed predictions generally outperforming calibration on raw probability scales [24]. Applying such procedures to KRR and ridge regression outputs could

substantially recover their operational utility in confidence-sensitive breed assignment contexts, without sacrificing their otherwise strong discriminative performance.

4.4 Overfitting in High-Dimensional Genomic Data: Sample Depth and Model Complexity

The pronounced species asymmetry in train-to-test generalisation gaps for boosting-family classifiers is interpretable within a broader framework of high-dimensional learning with limited samples. In cattle, per-breed sample sizes ranged from 17 to 31 individuals against the 5,000 autoencoder-selected features used for classification — a count still far exceeding the per-breed sample size — conditions that represent a particularly acute instance of what Berisha et al.[23] described as the "perfect storm" of high-dimensional data and small sample size, wherein complex models incur large blind spots and their test performance estimates are unreliable guides to deployment behaviour. In sheep, the substantially larger per-breed sample sizes (35–120) and lower feature-to-sample ratio moderated this effect, consistent with the observed compression of generalisation gaps across all model families.

4.5 Breed-Level Discrimination Difficulty as a Genomic Proximity Indicator

The identification of specific breeds that resist accurate classification across all or most models provides information that aggregate accuracy metrics do not. In cattle, Braunvieh was the sole breed for which no classifier achieved perfect recall, a result attributable to its genomic proximity to other European taurine breeds rather than to any model-specific limitation — as evidenced by the uniformity of the ceiling across 14 algorithmically distinct classifiers. In sheep, five commercially selected European-derived breeds produced shared errors under both top-ranked models — Australian Suffolk, Irish Suffolk, Australian Poll Dorset, Border Leicester, and German Texel — forming a consistent error cluster. Despite representing distinct breed groups, these breeds share exposure to intensive commercial selection pressures and, in the case of the Australian-derived breeds, partially overlapping founder populations, which may have generated convergent allele frequency profiles that reduce inter-breed genomic distance below the resolution threshold of standard SNP panels.

Moradi et al.[15] demonstrated in Iranian indigenous sheep breeds that discriminant SNP selection using linear methods could achieve complete separation with 201 informative markers, but noted that closely related breeds require a greater density of breed-informative

markers — consistent with the interpretation that the persistent misclassification of Suffolk-type sheep in this study reflects a genomic resolution problem rather than a modelling failure. Kumar et al.[25] showed that an ML-optimised panel of 23–48 SNPs achieved breed identification accuracies of 95–98% in Tharparkar cattle, illustrating that targeted marker panel design can address discrimination challenges that whole-chip approaches leave unresolved. For breeds like Braunvieh and Australian Suffolk, dedicated informative SNP panels or higher-density genomic data may be necessary to achieve the classification reliability expected in applied traceability contexts.

4.6 Statistical Equivalence and the Role of Calibration in Model Selection

The McNemar pairwise analysis revealed that, within the top-performing classifier cluster, individual-sample prediction errors were statistically indistinguishable, with all within-group adjusted p-values equal to 1.000 in cattle. This statistical equivalence has a practical implication that is not immediately apparent from accuracy tables: when discriminative performance is saturated, calibration quality becomes an important additional criterion for model selection. The Composite and Borda rankings formalised this principle — models with similar or identical accuracy diverged substantially in their weighted rankings once Brier score, log-loss, and ECE contributed their respective penalty weights. Random Forest, for instance, occupied a substantially lower Borda position than its accuracy rank would suggest, reflecting that its calibration-metric performance reduced its standing in a metric-equal weighting scheme. This pattern supports the view that accuracy alone is an insufficient basis for model selection in genomic classification tasks where the downstream application involves probability-based decisions, and that multi-metric evaluation frameworks of the type applied here are warranted even when headline discrimination metrics converge.

Thorsrud et al.[21] reached a related conclusion in the context of genomic breeding value prediction, noting that simpler models such as GBLUP may be preferred for operational use when their performance is statistically indistinguishable from more complex alternatives, given their lower optimisation burden. This study extends this reasoning by showing that within a statistically equivalent performance cluster, the models with the best-calibrated probability output — PyTorch MLP, logistic regression, and elastic net in the present results — provide the most complete operational profile.

4.7 Practical Implications for Genomic Breed Assignment

The results of this study suggest that genomic breed assignment in both cattle and sheep can be achieved with high accuracy and reliability using standard whole-chip SNP data, provided that model selection accounts for the full operational context rather than discriminative accuracy alone. For applications requiring only rank-based breed assignment, the choice among the statistically equivalent top-tier classifiers has limited practical consequence. For applications where probability outputs inform thresholded decisions — including traceability audits, certification of pure-bred status, or population management in conservation breeding programmes — PyTorch MLP, logistic regression, and elastic net offer the most dependable combination of accuracy and calibration in the present datasets. Marshall et al.[26] highlighted genomic breed assignment as transformational for livestock management contexts where phenotype-based identification is unreliable, and the present results indicate that even modest-cost linear classifiers can deliver the accuracy levels required for such applications.

A further practical consideration is the potential for panel reduction. The finding that autoencoder-based compression to a fraction of the original SNP count preserved discriminative performance, combined with published demonstrations that breed-informative SNP subsets of fewer than 1,000 markers can achieve comparable accuracy in other species [2] and that panels as small as 23–48 SNPs can suffice in targeted single-breed identification tasks [25], points toward the feasibility of reduced-cost genotyping solutions for routine breed assignment. Whether such reductions are applicable to multi-breed panels of the scale evaluated here — 20 breeds per species, spanning diverse geographic origins — remains to be determined by studies specifically designed to identify cross-breed informative subsets.

4.8 Limitations and Future Directions

Several limitations should be considered when interpreting the present results. First, all models were evaluated on held-out samples drawn from the same reference populations used for training; no external validation against independent populations, different genomic platforms, or admixed individuals was performed. In addition, the two species were genotyped on different arrays, so cross-species comparisons are confounded with platform ascertainment and cannot isolate biological from technical differences. Generalisation of the calibration and accuracy results to new populations, particularly those with different breed compositions or lower-quality genotype data, cannot be assumed. Second, per-breed sample sizes in cattle were modest for several breeds (n = 17–31), which may have inflated the generalisation gap for complex models and limited statistical power in breed-level recall estimates. Third, the KRR

implementation used here employed a one-vs-rest decomposition with softmax normalisation, which departs from the probabilistic framework of classical Bayesian GBLUP; its calibration results should not be read as properties of GBLUP. Future work could usefully examine whether post-hoc calibration of KRR and ridge regression outputs recovers their operational utility in confidence-sensitive contexts, and whether autoencoder-derived feature subsets retain their discriminative properties when applied to independent populations not represented in the reference panel.

5. Conclusions

These results bear on how genomic breed-assignment systems should be evaluated and deployed, rather than only on which algorithm scores highest. Once discrimination saturates, models that are statistically indistinguishable on accuracy can still differ substantially in the reliability of their probability outputs; calibration therefore warrants treatment as an evaluation criterion commensurate with discrimination wherever assignment decisions are conditioned on confidence thresholds. The close correspondence between low-dimensional population structure and classification difficulty further suggests that the principal constraint on performance is genomic resolution between proximate breeds rather than model capacity — a constraint more plausibly addressed through targeted breed-informative marker panels than through algorithm selection, and one that population-structure analysis can flag *a priori*. These inferences rest on within-sample cross-validation of reference populations genotyped on different arrays; external validation against independent, admixed, and platform-divergent samples is a necessary next step.

Acknowledgements

The authors thank the International Sheep Genomics Consortium (ISGC) and the contributors of the Dryad Digital Repository for making the genotype datasets publicly available.

Funding

This research received no external funding.

Author Contributions

Y.Y. conceptualized and designed the study, performed the genomic prediction and machine learning analyses, and wrote the manuscript. B.T. collected and curated the data.

Data Availability

Cattle genotype data were obtained from the Dryad Digital Repository (https://doi.org/10.5061/dryad.th092; WorldWideCattleDryad.zip), as originally published by Decker et al. (2014). Sheep genotype data were retrieved from the International Sheep Genomics Consortium (ISGC; https://www.sheephapmap.org/) [17]. All data used in this study are publicly available from the sources cited above.

Ethics Declarations

Competing interests

The authors declare no competing interests.

Ethical statement

This study was conducted exclusively using publicly available genotype datasets. No animal experiments were performed, and no ethical approval was required.

References

1. Wilmot, H. & Gengler, N. Good practice for assignment of breeds and populations—a review. *Front. Anim. Sci.* **6**, (2025).
2. Zhao, C. *et al.* Breed identification using breed-informative SNPs and machine learning based on whole genome sequence data and SNP chip data. *J Anim Sci Biotechnol* **14**, 85 (2023).
3. Wilmot, H., Niehoff, T., Soyeurt, H., Gengler, N. & Calus, M. P. L. The use of a genomic relationship matrix for breed assignment of cattle breeds: comparison and combination with a machine learning method. *J Anim Sci* **101**, skad172 (2023).

4. Burocziova, M. & Riha, J. Horse breed discrimination using machine learning methods. *J Appl Genet* **50**, 375–377 (2009).
5. Abdollahi-Arpanahi, R., Gianola, D. & Peñagaricano, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet Sel Evol* **52**, 12 (2020).
6. Liu, R. *et al.* Evaluation of six machine learning classification algorithms in pig breed identification using SNPs array data. *Animal Genetics* **54**, 113–122 (2023).
7. Manzoori, S., Farahani, A. H. K., Moradi, M. H. & Kazemi-Bonchenari, M. Detecting SNP markers discriminating horse breeds by deep learning. *Sci Rep* **13**, 11592 (2023).
8. Zhang, Z. *et al.* A deep learning strategy for accurate identification of purebred and hybrid pigs across SNP chips. *J Anim Sci Biotechnol* **16**, 116 (2025).
9. Su, R. *et al.* Genomic selection in pig breeding: comparative analysis of machine learning algorithms. *Genet Sel Evol* **57**, 13 (2025).
10. Miao, J. *et al.* A web tool for the global identification of pig breeds. *Genet Sel Evol* **55**, 18 (2023).
11. Zhao, Y., Liu, X., Liang, B. & Jiang, L. rPIMS: a ShinyR package for the precision identification and modelling of livestock breeds using genomic data and machine learning approaches. *Bioinformatics Advances* **5**, vbaf077 (2025).
12. Ryan, C. A., Berry, D. P., O'Brien, A., Pabiou, T. & Purfield, D. C. Evaluating the use of statistical and machine learning methods for estimating breed composition of purebred and crossbred animals in thirteen cattle breeds using genomic information. *Front. Genet.* **14**, (2023).
13. Gao, J. *et al.* Screening Discriminating SNPs for Chinese Indigenous Pig Breeds Identification Using a Random Forests Algorithm. *Genes (Basel)* **13**, 2207 (2022).

14. Li, D., Hu, X. & Peng, Y. The classification method of donkey breeds based on SNPs data and machine learning. *Front Genet* **16**, 1496246 (2025).
15. Moradi, M. H., Khaltabadi-Farahani, A. H., Khodaei-Motlagh, M., Kazemi-Bonchenari, M. & McEwan, J. Genome-wide selection of discriminant SNP markers for breed assignment in indigenous sheep breeds. *Annals of Animal Science* **21**, (2021).
16. Decker, J. E. *et al.* Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLOS Genetics* **10**, e1004254 (2014).
17. Kijas, J. W. *et al.* Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol* **10**, e1001258 (2012).
18. Ausmees, K. & Nettelblad, C. A deep learning framework for characterization of genotype data. *G3 Genes|Genomes|Genetics* **12**, jkac020 (2022).
19. López-Cortés, X. A., Matamala, F., Maldonado, C., Mora-Poblete, F. & Scapim, C. A. A Deep Learning Approach to Population Structure Inference in Inbred Lines of Maize. *Front. Genet.* **11**, (2020).
20. Farooq, M., van Dijk, A. D. J., Nijveen, H., Mansoor, S. & de Ridder, D. Genomic prediction in plants: opportunities for ensemble machine learning based approaches. *F1000Res* **11**, 802 (2022).
21. Thorsrud, J. A., Evans, K. M., Quigley, K. C., Srikanth, K. & Huson, H. J. Performance Comparison of Genomic Best Linear Unbiased Prediction and Four Machine Learning Models for Estimating Genomic Breeding Values in Working Dogs. *Animals* **15**, 408 (2025).
22. Dormann, C. F. Calibration of probability predictions from machine-learning and statistical models. *Global Ecology and Biogeography* **29**, 760–765 (2020).
23. Berisha, V. *et al.* Digital medicine and the curse of dimensionality. *NPJ Digit Med* **4**, 153 (2021).

24. Ojeda, F. M. *et al.* Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine* **42**, 5451–5478 (2023).
25. Kumar, H. *et al.* Machine Learning-Aided Ultra-Low-Density Single Nucleotide Polymorphism Panel Helps to Identify the Tharparkar Cattle Breed: Lessons for Digital Transformation in Livestock Genomics. *OMICS* **28**, 514–525 (2024).
26. Marshall, K. *et al.* Livestock Genomics for Developing Countries - African Examples in Practice. *Front Genet* **10**, 297 (2019).