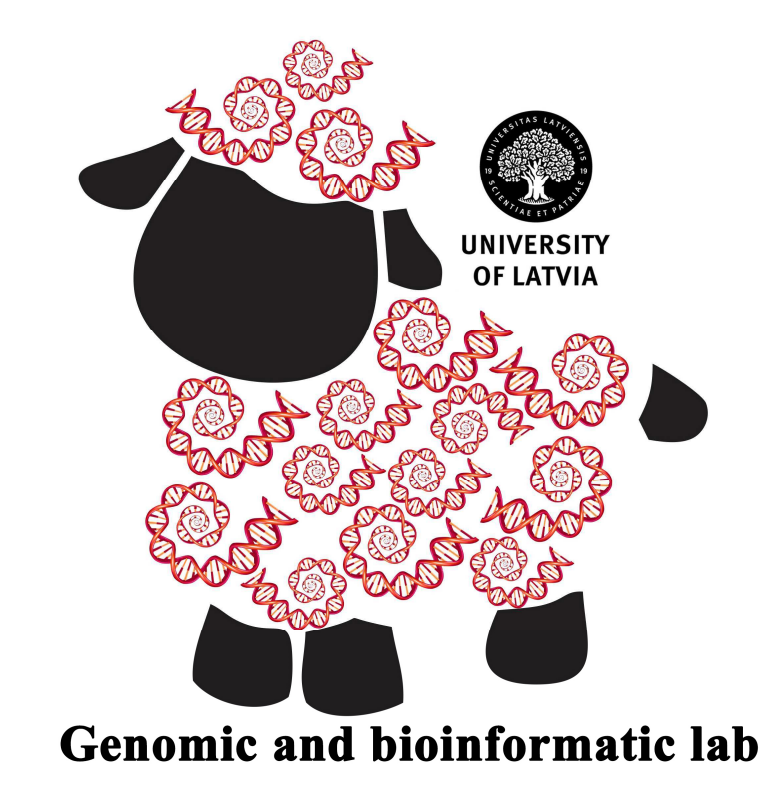


DEVELOPMENT OF A FRAMEWORK FOR ALGORITHMS TO PREDICT PHENOTYPIC TRAITS IN THE LATVIAN DARK-HEAD SHEEP BREED



M. Martins¹, I. Trapina¹, S. Plavina¹, D. Malakovska¹, N. Krasnevskā¹, J. Paramonovs¹, D. Kairisa^{2,3}, N. Paramonova¹

¹ Genomics and Bioinformatics, The Department of Pharmaceutical Sciences, Faculty of Medicine and Life Sciences, the University of Latvia, Riga, Latvia;
² Department of Animal Sciences, Faculty of Agriculture and Food Technology, Latvian University of Life Sciences and Technologies, Jelgava, Latvia;
³ Latvian Sheep Breeders' Association, «Klimpas», Jeru parish, Valmiera region, Latvia;

Introduction: marker-assisted algorithm for FCR trait prediction

Feed efficiency is one of the most **important economic indicators** in intensive sheep farming, as feed costs account for a significant portion of total livestock production costs. One of the most **commonly used** indicators of feed efficiency is the **Feed Conversion Ratio (FCR)**, which describes the ratio of feed consumed to live weight gain over a specific period of time. **Traditional breeding methods** may be **limited** in improving such traits, especially when **heritability is low** and the **factors** influencing **them are non-linearly** related. Therefore, increasing attention is being paid to the use of molecular markers in selection and **machine learning models**.

Biological samples and data

- All of the used sheep **data** was **obtained** from animals between **2022** and **2025**.
- In total there were **365** DNS sample collection from which **52%** was **LT (*Latvijas tumšgalves*; Latvian dark-head)** breed lambs.
- The **GeneSeek® Genomic Profiler™ Ovine 50K** (GGP Ovine 50K; NEOGEN, USA) was used to obtain **SNP data**.



The encountered FCR trait problems and solution

- Low heritability (5-25%)**
 - FCR is heavily **influenced** by the **environment**
 - FCR trait has very **scattered values**
- The **solution** was to allow the predicted models to have an **error interval** (10%, 15%, 20%) and to use **classification**, **splitting** the **data** into high, low, medium classes (using +/- 0.5 SD and +/- 1SD)

DATA after QC

Quality control filtering removed duplicate markers and SNPs with low minor allele frequency (MAF), resulting in a final dataset of **48,981** SNPs.

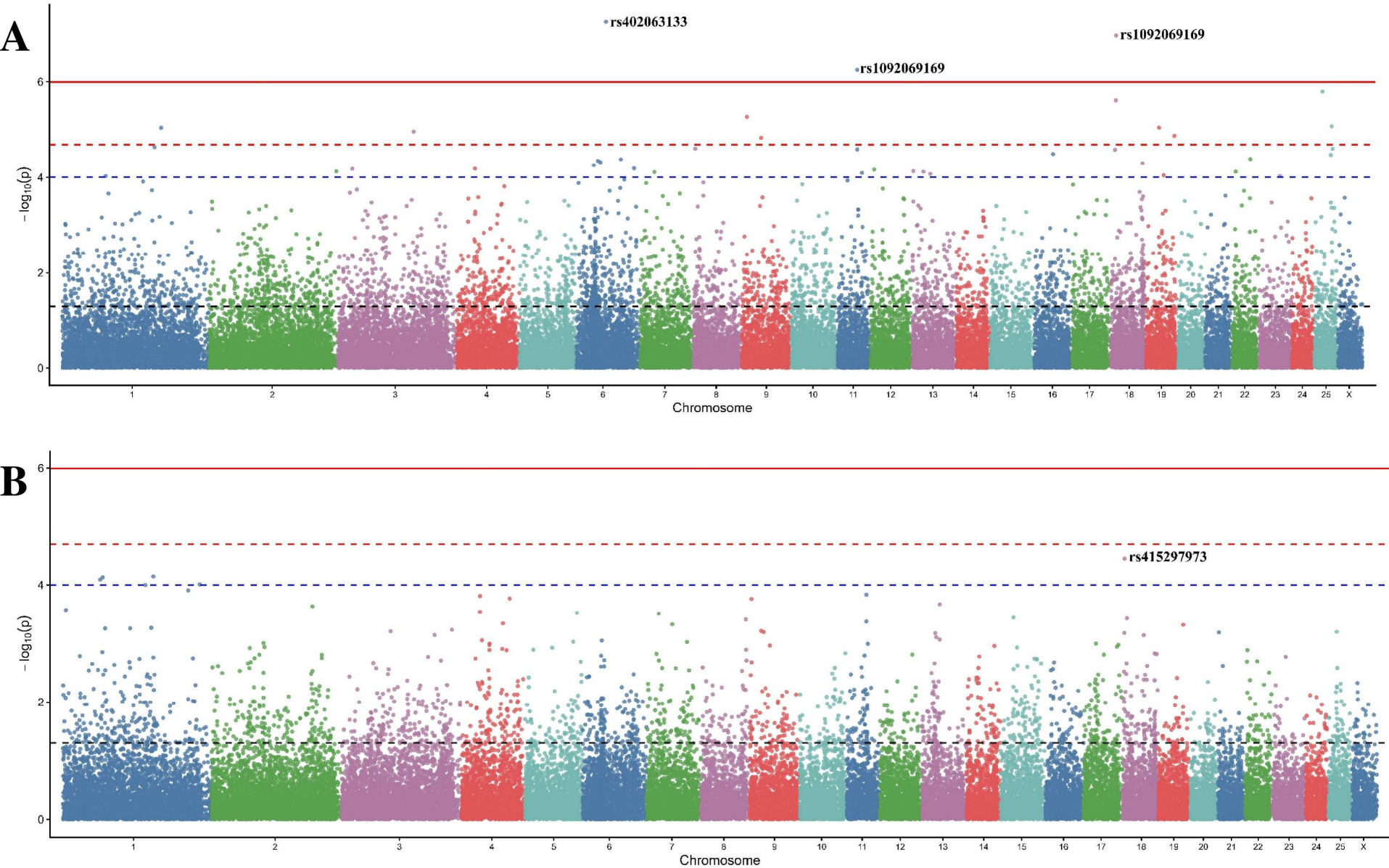


Methods

- Performed a **genome-wide association study (GWAS)** using a **linear mixed model** and a **genomic kinship matrix** to identify SNPs associated with feed conversion ratio (FCR).
- Developed a machine learning **pipeline** to evaluate different SNP subsets, parameter combinations, and model configurations.
- Applied **automated hyperparameter optimization** to identify the best-performing models.
- Performed **manual hyperparameter fine-tuning** of the top models to further improve predictive performance.

RESULTS

GWAS



- Whole-genome significance threshold: 5.99 (red line)
- Recommended significance threshold: 4.69 (red dotted line)
- Moderate association: 4 (blue dotted line)
- Weak association: 3

Limit -log10(p)	SNP LT	SNP ALL
5.99	0	3
4.69	0	10
4.00	5	31
3.00	53	179
1.30	2449	3854
Total SNP P>0.05	2507	4077

Regression

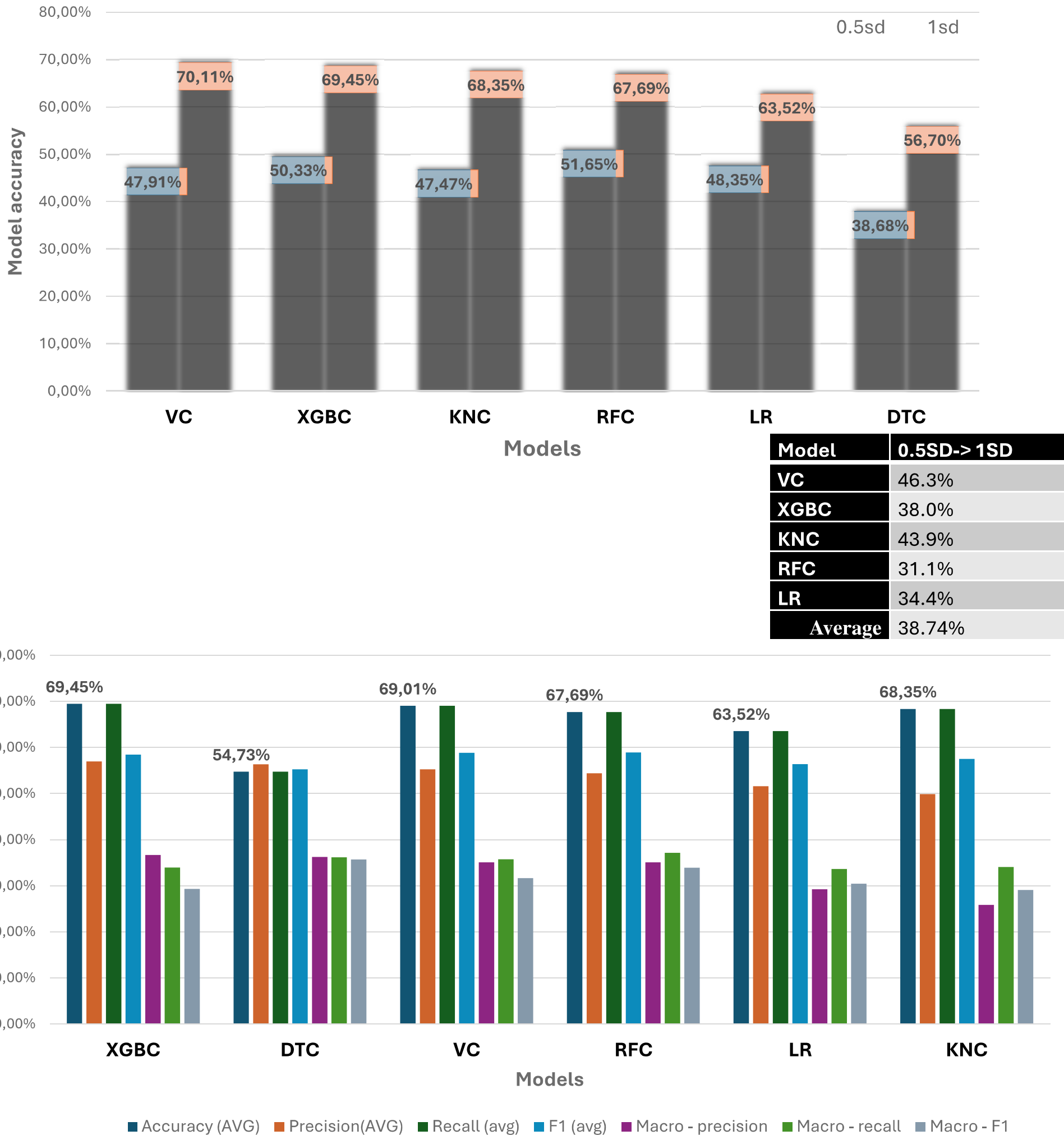
Model	Model parameters			Allowed error			Precision gained from increasing the allowed error	
	SNP	X	Y	10%	15%	20%	10%->15%	15%->20%
XGBR	2000	X_all	Original	54.06	70.77	79.56	30.9	12.40
VR	50	X_all	Original	52.75	70.33	80.00	33.3	13.70
SVR	50	X_snp	Original	54.95	70.33	79.12	28.0	12.50

- X_all – includes SNP and 2 phenotypic values
- X_snp includes only SNP values
- Y original indicates that the Y value matrix wasn't rounded up (which was also tested)

The increase in accuracy from 10% to 15% was around 33.6 for all of the models

- Models trained on **continuous** (non-rounded) FCR values **achieved**, on average, ~4.7% higher prediction (exception was DT in cases till 2000 SNPs).
- Models trained using only **LT** sheep achieved slightly lower prediction accuracies than models trained on the full population.
- The **highest** prediction accuracies were **generally** obtained using **50–2000 SNP markers**, while using all available SNPs did not consistently improve model performance.
- Hyperparameter **optimization** substantially **improved** model **performance**.

Classification



CONCLUSIONS

- Continuous** (non-rounded) FCR **values** consistently **improved** model **performance**, indicating that preserving phenotypic information is important for accurate prediction.
- Regression** models **outperformed classification** approaches, suggesting that FCR is better represented as a continuous trait rather than discrete classes.
- The **best** prediction results **were** achieved using **XGBoost, Support Vector Regression and Voting Regressor** models, highlighting their ability to capture complex non-linear relationships between SNP markers and FCR.
- The highest prediction accuracy (70.77%) was achieved by XGBoost using 2000 SNPs and phenotypic features.
- Models** trained on the **full population** achieved **higher accuracies** than models trained only on Latvian Dark-head sheep, likely due to larger sample size and stronger SNP associations.
- Future improvements** may be achieved by developing LT-specific SNP panels and incorporating **additional environmental and pedigree** information into the prediction models.