DOI 10.32900/2312-8402-2025-135-66-75
UDC 636.082.2:575:519.68

APPLICATION OF SUPPORT VECTOR REGRESSION (SVR) FOR IDENTIFICATION AND "REVERSE ENGINEERING" OF HIGH-PRODUCTIVITY ALLELIC COMBINATIONS IN CATTLE

Yuriy LIASHENKO, C.Sc. (Agr.), S.R.S.,
https://orcid.org/0000-0003-2747-476X
Livestock Farming Institute of NAAS of Ukraine, Kharkiv, Ukraine

Traditional selection in cattle heavily relies on linear mixed models (BLUP), which are effective but limited in modeling non-linear genetic interactions (epistasis). Machine learning (ML) algorithms offer an alternative capable of detecting complex dependencies in genetic data. The aim of this work was to test the Support Vector Regression (SVR) methodology for predicting milk productivity and to develop a "reverse engineering" approach to identify optimal allelic combinations based on a limited and heterogeneous set of genetic markers.

The study was conducted on a sample of 81 Ukrainian Red-and-White dairy cows. Genotypes for 3 QTLs (PRL, LEP, TNF-α) were used, which were transformed into 12 binary features (One-Hot encoding). Milk yield (305 days) and fat content (kg) were used as target variables for building the SVR model. The target variable (milk yield) was standardized using StandardScaler. The model was trained using 5-fold cross-validation with hyperparameter tuning (GridSearchCV), comparing both non-shuffled and shuffled data splits. A synthetic "solution space" (54 combinations) was generated to identify "ideal" genotypes, which was then analyzed by the trained SVR model.

Three-way ANOVA did not reveal a statistically significant ($p < 0.05$) effect of the main factors (PRL, LEP, TNF-α) or their interactions on milk yield, although PRL showed a borderline trend ($p=0.055$). SVR models trained on non-shuffled data failed, yielding negative R^2 values (down to -0.066), indicating overfitting. However, the model using all 3 markers (12 features) combined with 5-fold cross-validation with shuffling (shuffle=True) achieved the best, albeit practically negligible, positive result ($R^2 = 0.0064$) using a non-linear 'rbf' kernel, with an estimated RMSE of ~790 kg. The "reverse engineering" approach identified hypothetical complex genotypes (Top 3: CC-CC-AD, CT-CC-AD, CC-CC-AB) with a predicted yield (up to 5173 kg) significantly higher than the herd average (4838 kg).

The study confirmed the methodological suitability of SVR for analyzing heterogeneous genetic data and "reverse engineering" selection goals, even on a critically small sample (n=81). The low R^2 values highlight that the primary limitation is the small sample size relative to the number of features, which prevents the model from capturing reliable predictive signals. This approach serves as a powerful analytical complement to traditional BLUP methods, providing a framework for identifying desirable "genetic formulas" for targeted selection once larger datasets become available.

Keywords: machine learning (ML), support vector regression (SVR), prediction, genetic markers, dairy cattle productivity.

ЗАСТОСУВАННЯ МЕТОДУ ОПОРНИХ ВЕКТОРІВ (SVR) ДЛЯ ІДЕНТИФІКАЦІЇ ТА "ЗВОРОТНОГО ІНЖИНІРИНГУ" ВИСОКОПРОДУКТИВНИХ АЛЕЛЬНИХ КОМБІНАЦІЙ У ВРХ

Юрій ЛЯШЕНКО, к. с.-г. н., с. н. с., https://orcid.org0000-0003-2747-476X
Інститут тваринництва НААН, Харків

Традиційна селекція ВРХ спирається на лінійні змішані моделі (BLUP), ефективні, але обмежені у врахуванні нелінійних генетичних взаємодій. Алгоритми машинного навчання (МН) дають змогу виявляти складні залежності у генетичних даних. Мета роботи — апробувати регресію опорних векторів (SVR) для прогнозування молочної продуктивності та розробити підхід «зворотного інжинірингу» для визначення оптимальних алельних комбінацій на основі обмеженого гетерогенного набору маркерів. Дослідження виконано на вибірці з 81 тварини української червоно-рябої молочної породи. Використано генотипи трьох QTL (PRL, LEP, TNF-α), перетворені у 12 бінарних ознак (One-Hot encoding). Цільова змінна — надій за 305 днів. Дані поділено на тренувальний (80%) і тестовий (20%) набори, масштабування проведено за допомогою StandardScaler. Модель SVR навчали з 5-кратною перехресною валідацією та налаштуванням гіперпараметрів (GridSearchCV), порівнюючи перетасовані й неперетасовані дані. Для пошуку оптимальних генотипів створено синтетичний «простір рішень» (54 комбінації), проаналізований навченою моделлю SVR.

Трифакторний ANOVA не виявив статистично значущого впливу факторів (PRL, LEP, TNF-α) або їх взаємодій на надій ($p < 0,05$), хоча для PRL відзначено пограничну тенденцію ($p = 0,055$). Моделі SVR на неперетасованих даних показали від'ємні R^2 (до –0,066), що свідчить про перенавчання. Натомість модель з усіма трьома маркерами (12 ознак) і 5-кратною перехресною валідацією з перетасуванням (shuffle=True) досягла найкращого, хоч і незначного, результату ($R^2 = 0,0064$; RMSE ≈ 790 кг) з ядром «rbf». Підхід «зворотного інжинірингу» визначив потенційно оптимальні генотипи (Топ-3: CC-CC-AD, CT-CC-AD, CC-CC-AB) з прогнозованим надоєм до 5173 кг, що перевищує середній рівень по стаду (4838 кг).

Дослідження підтвердило методологічну придатність SVR для аналізу гетерогенних генетичних даних та цілей відбору за допомогою «зворотної інженерії», навіть на критично малій вибірці (n=81). Низькі значення R^2 підкреслюють, що основним обмеженням є малий розмір вибірки відносно кількості ознак, що заважає моделі фіксувати надійні прогностичні сигнали. Цей підхід служить потужним аналітичним доповненням до традиційних методів BLUP, забезпечуючи основу для визначення бажаних «генетичних формул» для цілеспрямованого відбору, коли стануть доступними більші набори даних.

Ключові слова: машинне навчання (МН), метод опорних векторів (SVR), прогнозування, генетичні маркери, молочна продуктивність корів

Introduction. The estimation of animal breeding values (EBV) forms the basis of modern selection. For decades, BLUP (Best Linear Unbiased Prediction) models and their genomic modifications (GBLUP, ssGBLUP) have remained the recognized standard in the field (Chafai et al., 2023). These methods are linear mixed models that effectively account for additive genetic variation using relationship (A) or genomic (G) matrices.

However, quantitative traits such as milk productivity are the result of complex biological processes involving non-additive genetic effects, particularly dominance and epistasis (gene-gene interactions) (Mackay TF, 2014). Linear BLUP models account for epistasis only indirectly, and its explicit inclusion in prediction remains a complex task (Alves K et al., 2023).

In parallel, alternative approaches based on machine learning (ML) algorithms are being developed, such as Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting Machines (GBM), and neural networks (Azodi et al., 2019, Mendoza et al., 2019). The key advantage of these methods is their ability to model complex, non-linear interactions directly from the data (Pérez-Enciso et al., 2019). This is particularly relevant for populations with limited or heterogeneous genomic information (e.g., when using data of different types – immunogenetic, microsatellite, SNP).

Unlike BLUP, ML models can serve as a complementary prediction system that focuses exclusively on the direct "marker → phenotype" relationship, eliminating the need for complex relationship matrices.

Current research extends beyond predicting the phenotype based on a known genotype ($G \rightarrow P$). Rather, of particular interest is the "reverse engineering" of genetic data (Rockman, 2008; Choi et al., 2024). This approach treats a trained machine learning (ML) model as a powerful computational simulator (Liu et al., 2025). It allows for *in silico* resolution of the inverse problem ($P \rightarrow G'$): "What allelic combination (G') is necessary to achieve a target productivity level (P)?" (Cavallaro et al., 2024)..

Today, thanks to the development of molecular technologies, a significant amount of data on the polymorphism of key QTL loci and their associations with economically valuable traits has been accumulated. This has become the basis for the active implementation of marker-assisted selection (MAS). For certain cattle breeds, comprehensive model genotypes, so-called "desirable genotype formulas," have already been developed, reflecting optimal allelic combinations for individual candidate genes (e.g., *CSN3, DGAT1, LEP, GH, PIT-1*, etc.) (Kopylov et al., 2015; Berezovskyi et al., 2015, Kopylov et al., 2016; Hladiy et al., 2018, Ivashchenko, 2023).

However, this approach has a fundamental limitation. In modern domestic and global practice, research mostly focuses on individual loci, while the analysis of complex genotypes (haplotypes) and their combined effect on productivity is almost never carried out. Instead, desirable genotypes of individual loci are usually mechanically combined into "formulas" that are actually based on an additive model—an assumption that the overall effect is a simple sum of the contributions of individual genes.

Such an approach deliberately ignores epistasis—complex non-linear interactions between different genes. Quantitative traits, including milk productivity, are polygenic in nature, and their phenotypic expression is determined not so much by the presence of individual "favorable" alleles as by their harmonious combination. The "optimal" allele of one gene may prove to be ineffective or even unfavorable in combination with a specific allele of another gene.

This is why the use of machine learning (ML) methods in this context is extremely promising. Traditional linear models have limitations, whereas ML algorithms like SVR or Random Forest can detect hidden non-linear relationships in complex datasets. They view the genotype as a comprehensive system, making it possible to account for epistatic effects. This enables a transition from the additive summation of markers to the identification of synergistic allelic combinations (haplotypes) that underlie the highest productivity.

The SVR algorithm is particularly attractive in this context. Unlike, for example, neural networks, which require massive amounts of data, SVR demonstrates high efficiency on relatively small samples (Vapnik, 1995). This is achieved through the "kernel trick," which allows the model to find non-linear dependencies in a highdimensional feature space without explicitly computing these complex interactions. Specifically, the Radial Basis Function (RBF) kernel, chosen for this study, is capable of modeling complex epistatic effects, making it an ideal candidate for analyzing polygenic traits on limited datasets (González-Recio et al., 2011).

The aim of this study was (1) to test the Support Vector Regression (SVR) methodology for predicting dairy cattle productivity and (2) to develop and test a "reverse engineering" approach to identify hypothetical high-productivity allelic combinations based on a limited set of heterogeneous genetic data.

Materials and methods. 1.*Sample and Phenotype Characteristics*. The study was conducted on a sample of cows (n=81) from the Ukrainian Red-and-White dairy breed population (State Enterprise Experimental Farm"Gontarivka," Kharkiv region). Data on milk yield for the first lactation (305 days) and fat content (kg) were used as the target traits (phenotypes). Descriptive statistics for the milk yield trait were: X_{Mean} =4838±91,5 kg, X_{Max}= 7093 kg, X_{Min}=2592 kg.

1. *Genotyping and Data Preparation*. Animals were genotyped using heterogeneous marker systems based on functional QTLs: prolactin (PRL, exon 4, 35106206C>T), leptin (LEP, exon 2, 73C>T), and tumor necrosis factor-alpha (TNF-α, exon 2, SSCP analysis).

To unify the data, the "One-Hot encoding" approach was applied. Each detected genotype (12 variants in total: PRL (3), LEP (3), TNF-α (6)) was converted into a separate binary feature (column), where 1 indicated the animal's possession of the allele and 0 indicated its absence. Thus, an initial feature matrix (X) with dimensions of 81 animals × 12 features (markers) was formed.

2. *Statistical Analysis (ANOVA)*. To preliminarily assess the associative links between individual genotypes (PRL, LEP, TNF) and milk yield ('Milk'), an analysis of variance (ANOVA) was performed. The analysis was carried out in the Python environment using the ols (Ordinary Least Squares) function from the statsmodels.formula.api library.

Both single-factor models (e.g., Milk ~ C(PRL)) and multi-factor models considering interactions (e.g., Milk ~ C(PRL) * C(LEP) and Milk ~ C(PRL) * C(LEP) * C(TNF)) were evaluated. The statistical significance of the effect of each factor and their interactions was determined using the F-test (using ANOVA Type II).

3. *SVR Modeling*. The Python programming language and the Scikit-learn library were used for the analysis (Pedregosa et al., 2011).

1) *Sample Formation:* The feature matrix (X, 81×12) and the target variable vector (y, 81×1, milk yield) were split into training (80% of data) and testing (20% of data) sets.

2) *Data Scaling:* As SVR is sensitive to scaling, the target variable (Y, milk yield) was standardized (StandardScaler) to a zero mean and unit variance. The feature matrix (X), consisting of binary data (0/1) after encoding, was not scaled.

3) *Model Training*: A Support Vector Regression (SVR) model was used. To find non-linear dependencies, the Radial Basis Function (RBF) kernel (kernel='rbf') was chosen, and the standard linear kernel (kernel='linear') was also tested for comparison.

Model optimization was conducted for two key hyperparameters:

- C (Regularization parameter): Controls the trade-off between minimizing the error on the training data and maximizing the margin. Low C values allow for a larger error (a softer margin), which prevents overfitting, whereas high C values attempt to minimize the error, risking overfitting.

- `epsilon` (ε-insensitive zone): Defines the width of the "tube" within which prediction errors are not penalized. This parameter allows the model to ignore minor "noise" in the data.

Optimal hyperparameters (C, epsilon, and kernel type) were automatically selected using an exhaustive Grid Search (`GridSearchCV`) with 5-fold cross-validation on the training set.

4) *Accuracy Assessment:* The predictive ability of the final model was evaluated on the test set (20% of data the model had not seen during training) using the coefficient of determination (R^2) and the Root Mean Squared Error (RMSE).

- R^2 (Coefficient of Determination): A metric showing the proportion (from 0 to 1) of the variance in the target variable that the model can explain. It is particularly important that R^2 can be negative ($R^2 < 0$) if the model performs worse than simple data averaging, which is a clear indicator of model non-viability.
- RMSE (Root Mean Squared Error): An absolute error measure expressed in the units of the target variable (in our case, kg of milk). It shows how much, on average, the model's predictions deviate from the actual values and is more interpretable than R^2 for assessing the practical magnitude of the error.

5 "Reverse Engineering" Methodology. After training and validation, the SVR model was used for "reverse engineering."

1) *Solution Space Creation:* Based on the 12 selected markers, a complete solution space was generated—a synthetic matrix `X_synth` containing all possible unique combinations of these alleles. The total volume of hypothetical genotypes was 54 (3×3×6) combinations.

2) *Hypothesis Prediction:* The trained SVR model (`model.predict(X_synth)`) was applied to each of the 54 hypothetical combinations to predict the *scaled* milk yield.

3) *Inverse Transformation:* The resulting predictions were returned to their original scale (kg) using an inverse transformation (`scaler_y.inverse_transform`), allowing absolute milk yield values to be obtained.

4) *Identification*: The obtained predictions were sorted in descending order to identify the TOP-10 best allelic combinations corresponding to the maximum predicted productivity.

Data processing and machine learning model construction were performed using the Google Colaboratory (Colab) cloud service.

Research results. Prior to building predictive machine learning models, an analysis of variance (ANOVA) was conducted to assess the presence of statistically significant associations between the genotypes of individual loci (PRL, LEP, TNF-α) and the target trait – milk yield for 305 days in the first lactation.

To evaluate not only individual effects but also possible epistatic interactions between loci, a three-way analysis of variance was performed. The analysis results (Table 1) indicate the absence of a statistically significant ($p < 0.05$) effect on milk yield for any of the three markers taken individually, as well as for factor interactions. The prolactin (*PRL*) locus showed a trend towards an association (p=0.055), but it did not reach the established significance level. The absence of strong linear associations indicates the difficulty of predicting this trait and justifies the need to apply more flexible non-linear machine learning methods like SVR.

For building the predictive models, SVR was used. The genotypes were transformed using One-Hot Encoding, resulting in a model with 6 features for 2 markers (*PRL+LEP*) and 12 features for 3 markers (*PRL+LEP+TNF-α*).

Table 1

Results of the three-way analysis of variance (ANOVA) of the influence of genotypes (*PRL, LEP, TNF-α*) on the milk yield of Ukrainian Red-and-White dairy cattle

Source of Variation	df	SS	MS	F	p
C(*PRL*)	2.0	5.287718e+06	2.64e+06	3.814	0.055
C(*LEP*)	2.0	2.803068e+05	1.40e+05	0.202	0.654
C(*TNF*)	5.0	4.461275e+06	8.92e+05	1.038	0.362
C(*PRL*):C(*LEP*)	4.0	5.585413e+05	1.40e+05	0.163	0.689
C(*PRL*):C(*TNF*)	10.0	3.024966e+06	3.02e+05	0.352	0.878
C(*LEP*):C(*TNF*)	10.0	5.055151e+06	5.06e+05	0.588	0.709
C(*PRL*):C(*LEP*):C(*TNF*)	20.0	7.549629e+06	3.77e+05	0.440	0.930
Residual	46.0	3.952052e+07	NaN	NaN	NaN

Notes. Residual - residuals, SS - sum of squares, MS - mean square, F - F-statistic, p - significance level.

Due to the extremely limited sample size (n=81), a 5-fold cross-validation method was used for objective model quality assessment. This method involves dividing the 81 observations into 5 equal parts ("folds"), approx. 16-17 animals each. The process is repeated 5 times: the model is trained on 4 parts (≈65 animals) and checks its accuracy (calculates R^2 and RMSE) on the 1 part (≈16 animals) it has "not seen." The final R^2 (or RMSE) score is the averaged value of these 5 separate tests. This provides a much more reliable estimate than a single 80/20 split.

Furthermore, we compared two cross-validation approaches:

1. Without shuffling (`shuffle=False`): Data is split sequentially. This can lead to a biased estimate if the data has a hidden order.
2. With shuffling (`shuffle=True`): Data is randomly permuted before being split into 5 folds. This creates more "truthful" and representative folds.

The impact of the cross-validation method proved critical for correct model evaluation. As shown below, models tested without shuffling (`shuffle=False`) demonstrated worse and less stable R^2 results than models tested with shuffling (`shuffle=True`), which provided more representative test folds.

Scenario 1: SVR Model (*PRL + LEP*)

When using the SVR model with two markers (*PRL* and *LEP*, 6 input features), the cross-validation results were unsatisfactory (Table 2).

Table 2

Comparison of SVR results for Scenario 1 (*PRL+LEP*)

Validation Method	Best R^2	RMSE (kg)	Best Hyperparameters
No Shuffle (shuffle=False)	-0.066	≈809	{'C': 0.1, 'epsilon': 0.01, 'kernel': 'rbf'}
With Shuffle (shuffle=True)	0.0037	≈793	{'C': 0.1, 'epsilon': 0.5, 'kernel': 'linear'}

In the first case (no shuffle), the R^2 is negative, indicating complete model non-viability. This is likely because the sequential data split led to unrepresentative test samples. Applying shuffling (`shuffle=True`) stabilized the evaluation, allowing a weak but positive R^2 = 0.0037 (0.4% of variance explained) to be obtained. Interestingly, `GridSearchCV` selected the `kernel='linear'` as optimal here, suggesting the model could not find useful non-linear dependency and the simple linear combination was slightly more robust against overfitting.

Scenario 2: SVR Model (*PRL + LEP + TNF-α*)

Adding the third marker (*TNF-α*), which increased the input features to 12, also demonstrated the benefits of shuffling (Table 3).

Table 3

Comparison of SVR results for Scenario 2 (*PRL+LEP+TNF-α*)

Validation Method	Best R²	RMSE (kg)	Best Hyperparameters
No Shuffle (shuffle=False)	-0.032	≈795	{'C': 0.1, 'epsilon': 0.01, 'kernel': 'linear'}
With Shuffle (shuffle=True)	0.0064	≈790	{'C': 0.1, 'epsilon': 0.01, 'kernel': 'rbf'}

Similar to scenario 1, the non-shuffled model showed a negative R² (-0.032). However, when applying shuffling, the model achieved the best score of all tested variants: R² = 0.0064.

It is important to note that in this case (with 12 features), `GridSearchCV` selected the non-linear `kernel='rbf'`. This suggests that adding the TNF-α marker allowed the model to capture an extremely weak, but statistically fixed, non-linear interaction between the markers, which neither the 2-factor model nor the classic ANOVA could detect.

Based on this best (though still very weak) SVR model (Scenario 2, shuffled, R²=0.0064), 54 hypothetical genotype combinations (3 *PRL* × 3 *LEP* × 6 *TNF-α*) were generated, and milk yield was predicted for each. The top 10 combinations with the highest predicted yield are presented in Table 4.

Table 4

The best hypothetical complex genotypes according to SVR prediction (R²=0.0064)

№	*PRL*	*LEP*	*TNF-α*	Milk yield (305 days), kg
1	CC	CC	AD	5173.2
2	CT	CC	AD	5139.7
3	CC	CC	AB	5057.8
4	CT	CC	AB	5035.9
5	CC	CT	AD	5022.0
6	CC	CC	AA	5013.6
7	TT	CC	AD	4988.0
8	CC	CT	AB	4964.9
9	CT	CC	AA	4940.0
10	CC	CC	AC	4929.1

The analysis shows that the model favors combinations containing the *PRL*-CC genotype (6 of the top 10), *LEP*-CC (8 of the top 10), and *TNF*-AD (4 of the top 10).

Discussion. The SVR modeling results are fully consistent with the preliminary ANOVA data (Table 1). ANOVA did not reveal a statistically significant linear relationship between the studied markers and milk yield, which explains why the predictive models built on the same data demonstrated an extremely low coefficient of determination (R^2). In essence, the machine learning models confirmed the ANOVA conclusion: in this sample (n=81), there is no strong predictive signal associated with the *PRL, LEP,* and *TNF-α* loci.

The study demonstrated the successful application of the SVR algorithm for working with heterogeneous genetic data (QTLs) under small sample size conditions. The key result is not so much the absolute prediction accuracy (which is expectedly low due to the N=81 to p=12 ratio) but the successful implementation of the "reverse engineering" methodology. Unlike traditional estimation, which provides a forecast for existing animals, our approach allowed us to generate an "ideal genetic formula" – a selection goal that can be pursued through targeted fixation of desired alleles.

The use of the non-linear `'rbf'` kernel in SVR potentially allowed the model to account for epistatic interactions between the 12 selected markers. It is likely that the high predicted yield of the TOP genotypes (Table 4) is due precisely to a successful combination of alleles, and not just their additive sum effect, which BLUP would estimate.

The study also highlighted the key challenges of working with small-sample biological data. Although classic ANOVA found no significant associations (p>0.05), applying SVR with shuffling (`shuffle=True`) allowed for a minimal positive $R^2 = 0.0064$ (Scenario 2). This indicates that adding the *TNF-α* marker and using the `'rbf'` kernel allowed the model to capture an extremely weak non-linear dependency that ANOVA could not detect.

However, an $R^2 = 0.0064$ means the model explains only about 0.6% (not 1%) of the milk yield variance. The calculated RMSE for this model was ~790 kg (with a data standard deviation σ=818 kg). This indicates that, despite the positive R^2, the model has no practical predictive value for selection.

The main reason for the low efficiency is the insufficient sample size (n=81). When we encode 3 markers, we get 12 input features. Training a model on 12 features with only 81 observations leads to the "curse of dimensionality" and severe overfitting, which we observed as negative R^2 values in most scenarios.

To improve the prediction, it is necessary to:

1. Significantly increase the sample size (to several hundred or thousand animals).
2. Include additional markers that have a proven strong association with the trait.
3. Include non-genetic factors (parity, calving season, feeding level) as additional features in the model.

Furthermore, it is worth noting that although Random Forest is often considered a more powerful algorithm, in parallel experiments (data not shown) on this same sample, it showed significantly worse results ($R^2 \approx$ -0.10 to -0.23). This indicates its stronger tendency to overfit on such a feature-to-observation ratio (12 to 81). This underscores the advantage of SVR (especially with regularization-promoting parameters like `C=0.1`) specifically in critically small sample conditions.

This study is primarily an important validation of the methodology for applying machine learning (SVR) for analyzing genetic data in animal husbandry. It clearly demonstrated the importance of correct model evaluation, particularly the use of cross-validation with shuffling (`shuffle=True`), to obtain reliable results on small samples.

It was shown that a negative R^2 (as in the non-shuffled scenario) is just as important a result as a positive one, as it clearly indicates the model's non-viability and prevents false conclusions that could be drawn from an inflated R^2 on the training set.

Although the current models have no practical predictive value, the developed algorithm (feature encoding, hyperparameter tuning, R^2/RMSE evaluation) provides a basis for further analysis on expanded datasets. The ML approach should be considered a powerful analytical tool that complements the traditional BLUP system. It provides a new methodological basis for identifying rare but high-productivity allelic combinations and for the operational management of the genetic fund.

Conclusions:

1. The Support Vector Regression (SVR) methodology confirmed its suitability for processing and modeling heterogeneous genetic data.
2. Multi-factor analysis of variance (ANOVA) did not reveal a statistically significant ($p < 0.05$) influence of the genotypes of the *PRL*, *LEP*, *TNF-α* loci, or their interactions, on the milk yield level in the studied sample of 81 animals.
3. Building SVR predictive models on a small number of observations (n=81) showed low efficiency. Models tested without data shuffling (`shuffle=False`) demonstrated negative R^2 (down to -0.066), indicating overfitting and lower accuracy than a simple mean model.
4. Applying cross-validation with shuffling (`shuffle=True`) stabilized the SVR model evaluation (*PRL+LEP+TNF-α*), allowing the best (though practically insignificant) result of $R^2 = 0.0064$ to be achieved with a non-linear `'rbf'` kernel.
5. The proposed "reverse engineering" approach successfully identified hypothetical genotypes with high predicted productivity; however, the best model's average error (RMSE) is ~790 kg, confirming its low predictive value for practical selection.
6. The insufficient sample size (n=81) relative to the number of input features (n=12) is the main limiting factor for building a reliable predictive model.
7. Further research requires a substantial increase in sample size and the testing of more complex ML algorithms (Random Forest, XGBoost) to improve prediction accuracy.

References

Alves K, Brito LF, Schenkel FS. (2023). Genomic prediction of fertility and calving traits in Holstein cattle based on models including epistatic genetic effects. *J Anim Breed Genet.* Sep;140(5):568-581. https://doi.org/10.1111/jbg.12810.

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH. (2019). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3 (Bethesda).* Nov 5;9(11):3691-3702. https://doi.org/10.1534/g3.119.400498.

Berezovskyi, O. V., Yu. P. Polupan, S. Yu. Ruban, & Kopylov K. V. (2015). Zv’iazok polimorfizmu za henamy κ-CN, TG5, LEP z molochnoiu produktyvnistiu koriv ukrains-kykh molochnykh pored [The connection of polymorphism to the κ-CN, TG5, LEP genes with the milk yield of cows of Ukrainian breeds]. Rozvedennya i henetyka tvaryn: mizhdvidomchyy tematychnyy zbirnyk –Animal Breeding and Genetics: interdepartmental thematic scientific collection. Kyiv, 49:154-164 (in Ukrainian)

Bergstra J., Komer B., Eliasmith C. et al. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. Comput. Sci. Discov. 8:014008. https://doi.org/10.1088/1749-4699/8/1/014008.

Cavallaro,C Cutello, V, Pavone, M & Zito, F. (2024). Machine Learning and Genetic Algorithms: A case study on image reconstruction. Knowledge-Based Systems 284 (111194). https://doi.org/10.1016/j.knosys.2023.111194.

Chafai, N., Hayah, I., Houaga, I., Badaoui, B. (2023). A review of machine learning models applied to genomic prediction in animal breeding. *Frontiers in Genetics*, 14:1150596. https://doi.org/10.3389/fgene.2023.1150596.

González-Recio, O., Forni, S. (2011). Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genet Sel Evol*, 43, 7. https://doi.org/10.1186/1297-9686-43-7.

Hladiy, M. V., Polupan, Y. P., Kovtun, S. I., Kuzebnij, S. V., Vyshnevskiy, L. V., Kopylov, K. V., & ShcherbakO. V. (2018). Scientific and organizational aspects of generation, genetics, reproduction biotechnology and protection of the genofonds in livestock breeding. Animal Breeding and Genetics, 56, 5-14. https://doi.org/10.31073/abg.56.01

Ivashchenko O. Yu. (2023). Henetychne riznomanittia populiatsii velykoi rohatoi khudoby za asotsiiovanymy z rezystentnistiu DNK-markeramy [Genetic diversity of cattle populations by resistance-associated DNA markers]: avtoreferat dys. ... d.filosof : 204 / O. Yu. Ivashchenko. — B.m., https://nubip.edu.ua/sites/default/files/u145/dis_ivashchenko.pdf (in Ukrainian).

Junhwa Choi, Sunghyun Cho, Subin Choi, Myunghee Jung, Yu-jin Lim, Eunchae Lee, Jaewon Lim, Han Yong Park & Younhee Shin. (2024). Genotype-Driven Phenotype Prediction in Onion Breeding: Machine Learning Models for Enhanced Bulb Weight Selection. Agriculture 14, 2239. https://doi.org/10.3390/agriculture14122239.

Kopylov, K. V., O. D. Biriukova, O. V. Berezovskyi, & Basovskyi D. M. (2015). Henetychnyi monitorynh v stadi ukrainskoi chervono-riaboi molochnoi porody za kompleksom heniv [Genetic monitoring in a herd of Ukrainian red-billed milk breed in a complex of genes]. Tekhnolohiya vyrobnytstva i pererobky produktsiyi tvarynnytstva - Technology o f production and processing of livestock products. Bila Tserkva. 1(116):28-31 (in Ukrainian).

Kopylov, K. V., O. I. Metlytska, N. B. Mokhnachova, & Suprovych T. M. (2016). Molekuliarno-henetychnyi monitorynh v systemi zberezhennia henetychnykh resursiv tvaryn [Molecular genetic monitoring in the system of conservation of genetic resources of animals]. Visnyk ahrarnoi nauky- Bulletin o f Agricultural Science. 6:43-47 (in Ukrainian)

Liu, M., Gao, Z., Chang, H., Li, S. Z., Shan, S., & Chen, X. (2025). G2PDiffusion: Cross-Species Genotype-to-Phenotype Prediction via Evolutionary Diffusion. arXiv preprint arXiv:2502.04684.

Mackay, T. F. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet.* Jan; 15(1):22-33. https://doi.org/10.1038/nrg3627.

Mendoza H., Klein A., Feurer M. et al. (2019). Towards automatically tuned deep neural networks. In: Hutter F. et al. (eds) Automated Machine Learning. Springer, Cham, pp. 135–149.

Pedregosa F, Varoquaux G, Gramfort A et al (2011). Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830.

Pérez-Enciso, M.; Zingaretti, L.M. (2019). A Guide on Deep Learning for Complex Trait Genomic Prediction. *Genes*, 10, 553. https://doi.org/10.3390/genes10070553.

Rockman, M. (2008). Reverse engineering the genotype–phenotype map with natural genetic variation. Nature 456, 738–744. https://doi.org/10.1038/nature07633.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer, New York. http://dx.doi.org/10.1007/978-1-4757-2440-0.

Wang, X., Shi, S., Wang, G. et al. (2022). Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. *Journal of Animal Science and Biotechnology*, 13:60. https://doi.org/10.1186/s40104-022-00708-0.