Capstone Project :

Predicting milk yield from cows using genetic markers

**By Kay Anantanawat**

**Date 11/05/24**

Table of Contents

(McHugh, 2013)Problem statement 3

Industry/ domain 3

Stakeholders 3

Business question 3

Data question 4

Data 4

Data science process 4

Data analysis 4

Cleaning steps 5

Modelling 5

Model 1: Multi Layer Perceptrons 5

Model 2: CNN 6

Model 3: CNN with regularisation L1 6

Model 4: CNN with regularisation L1 and L2 7

Outcomes 8

Implementation 8

Response to stakeholders 9

End-to-end solution 9

Bibliography 10

# Problem statement

Understanding how genes influence the physical characteristics of organisms has been a cornerstone of genetics since the time of Mendel. While some traits, like earlobe type or freckles, are controlled by single genes, others, such as fingerprint patterns or thumb dominance, are influenced by multiple genes. Complicating matters further, quantitative traits like height and skin color are controlled by a complex interplay of multiple genes. Despite this knowledge, accurately predicting physical traits using genetic information remains challenging.

Genomic prediction, a method that utilizes an organism's genetic data to forecast its traits, offers a promising avenue for tackling this challenge. Single-nucleotide polymorphisms (SNPs), representing changes in individual DNA bases, are abundant throughout an organism's genome and have significant effects on its physical characteristics. Leveraging SNPs data enables the prediction of various traits and has revolutionized breeding practices in agriculture, enhancing productivity, disease resistance, sustainability, and adaptability to changing conditions.

Current genetic prediction models, such as genomic best linear unbiased prediction (GBLUP) and Bayesian linear models, have proven effective in livestock and plant breeding (Zhang, et al., 2015). However, these models have limitations—they primarily consider additive effects of single genetic markers, assuming uniform contributions to traits. To address this, recent studies propose the use of Local Convolutional Neural Networks (CNNs) for genomic prediction, particularly for quantitative traits (Pook, Freudenthal, Korte, & Simianer, 2020). While GBLUP struggles to accurately predict quantitative traits, CNNs offer promising results. This project aims to explore the application of CNNs to SNPs data for predicting quantitative traits, potentially overcoming the limitations of traditional genetic prediction models and advancing our understanding of genetic influence on complex traits.

# Industry/ domain

The solution holds significant potential for two key industries: agriculture and healthcare. In agriculture, the model offers invaluable predictive capabilities for optimizing livestock breeding and enhancing the cultivation of economically valuable crops. By leveraging the model, farmers can forecast desirable trait values in animals and plants, leading to improved productivity and profitability.

Similarly, in the realm of healthcare, the model's utility extends to the realm of human genetics. Just as genetic data informs breeding decisions in agriculture, it also provides crucial insights into human health. The model's adaptability enables its application in identifying potential genetic diseases and predispositions in individuals. By analyzing genetic markers, healthcare professionals can preemptively address health concerns and tailor personalized treatment plans, ultimately improving patient outcomes and well-being.

# Stakeholders

The stakeholders encompass vital sectors in both agriculture and healthcare, comprising farmers, breeders, genetic counselors, and professionals and researchers in human health.

# Business question

How to improve the use of genetic data to predict the milk production in the cow to increase the efficiency of the business?

# Data question

Using Convolutional Neural Network (CNN), is it possible to use the SNPs data alone to predict the milk yield from Holstine cows?

# Data

The study utilizes genetic data sourced from a renowned provider for animal husbandry and breeding, Vereinigte Informationssysteme Tierhaltung w.V., based in Verden, Germany (Zhang, et al., 2015).This data pertains specifically to the German Holstein prediction population—a notable dairy cow breed recognized worldwide, including in Australia, where approximately 85% of dairy cows are Holsteins (Dairy Australia, n.d.). Known for their prolific milk production potential, Holstein cows can yield up to 10,000 litters of milk annually, equivalent to 14 cartons of 2L milk daily.

Comprising 5,024 bulls, the dataset features SNP data collected via DNA sequencing technology, utilizing the Illumina Bovine SNP50 Beadchip. Following rigorous quality control measures, 42,551 SNPs remain available for further analysis. Each bull's SNP data is represented by values of 0, 1, or 2, denoting the presence of alternative forms of the SNP. Specifically, 0 signifies no deviation from the reference bull in both SNP copies, while 1 indicates a single differing copy, and 2 signifies variance in both copies.

Augmenting this genetic data are highly reliable conventional estimated breeding values (EBVs)—quantitative estimates of an individual's genetic merit for a given trait—derived from the bulls. A positive EBV indicates preferable genetic merit compared to a reference individual, while a negative EBV signifies otherwise. Among the traits considered, such as milk fat percentage and somatic cell scores, milk yield emerges as the primary focus for this project. It's crucial to note that while the milk yield trait influences the milk production potential of female offspring, bulls themselves do not produce milk.

# Data science process

## Data analysis

The data pipeline is as shown in Figure 1.

Figure 1. Data pipeline. Data cleaning removing

The data consists of a data frame containing 42,552 columns (SNPs; features) and 5,024 rows (bulls) where the first column is the ID of the bull, and the rest of columns are the number of alternative alleles in each SNPs, value 0,1 and 2.

For a measure of quantitative trait, a separate data frame containing 5,024 rows (bulls) and 4 columns where the first column is the ID of the bull, and the other 3 columns are qualitative traits measured in the bull: milk yield, milk fat percentage and somatic cell score. For this project, milk yield is used as a target quantitative trait.

A graph of milk yield

Description automatically generated

Figure 2. Histogram of the milk yield for all 5,024 bulls.

The distribution of EBVs for the milk yield follows a normal distribution.

### Cleaning steps

First, the number of SNPs are reduced by removing SNPs that appear in the same form across 80% of the bull population. If 80% of the bull has the same number of genetic forms, either 0,1 or 2, the SNP is assumed to not contribute much to the genetic trait since they are shared by most of the bull population and are removed from the analysis. This results in 7,760 SNPs being removed, 34,792 SNPs remained.

Then, each SNPs was tested for independency. Considering that the value of each SNP is a categorical data, each SNP is tested for dependency to the other SNP using chi-square test (McHugh, 2013). The alpha value is 0.05. For each pair of SNPs, if the calculated p-value from chi-square test is equal or less than 0.5, the SNPs are not independent. One of the SNPs is being removed from the analysis. This results in 30,635 SNPs being removed, 4,157 SNPs remained.

After cleaning, the dataset is split into the training dataset (80%) and the testing dataset (20%).

## Modelling

The aim of the model is to predict the target variable, milk yield, based on SNPs data. The features used for training are 4,157 SNPs data that are independent of each other and has a higher diversity.

I made 4 different models: one MLP and 3 different CNN adapted from Pook et al. (2020).

### Model 1: Multi Layer Perceptrons

A comparison of blue dots

Description automatically generated with medium confidence

Figure 3. The plot showing predicted milk yield from Multiple Layer Perceptrons (2 fully connected layers) model VS true milk yield.

### Model 2: CNN

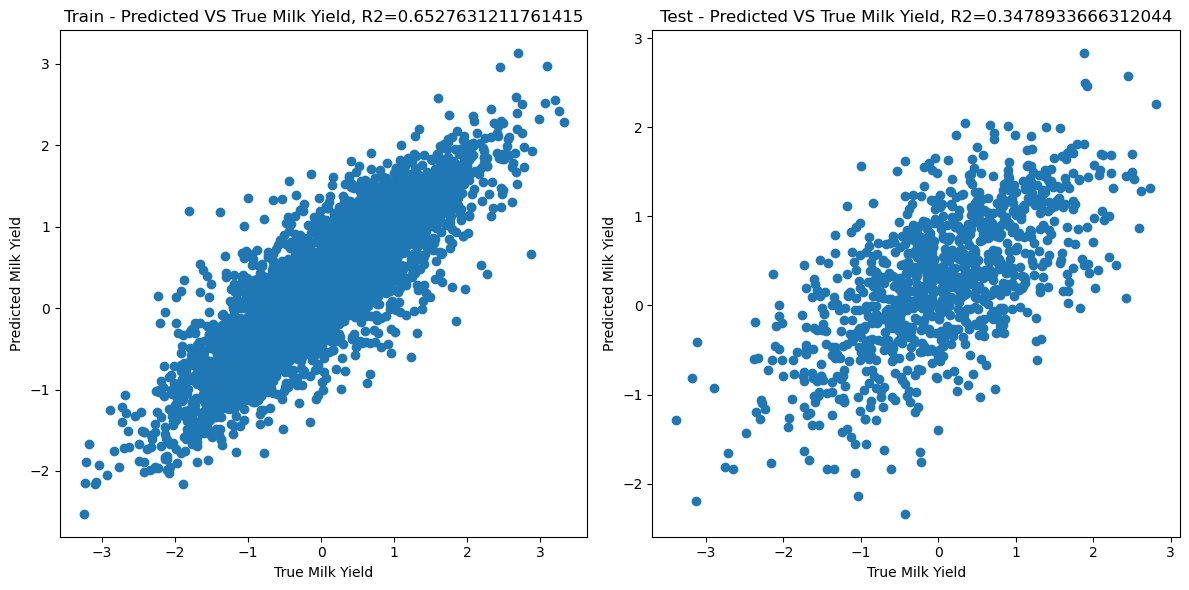


Figure 4. The plot showing predicted milk yield from CNN (2 fully connected layers) model VS true milk yield.

### Model 3: CNN with regularisation L1

A graph of blue dots

Description automatically generated with medium confidence

Figure 5. The plot showing predicted milk yield from CNN (2 fully connected layers) model with regularizer L1 VS true milk yield.

### Model 4: CNN with regularisation L1 and L2

A comparison of blue dots

Description automatically generated

Figure 6. The plot showing predicted milk yield from CNN (2 fully connected layers) model with regularizer L1 and L2 VS true milk yield.



Figure 7. Comparing of R-square scores between the training and the testing dataset from 4 different models: MLP, CNN, CNN with regularizer L1 and CNN with regularizer L1 and L2.

## Outcomes

After exhaustive experimentation and evaluation, I have determined that the Convolutional Neural Network (CNN) architecture, combined with both L1 and L2 regularization techniques, yields the highest performance metrics among the models tested. Despite this optimization effort, the analysis reveals that the model remains unable to fully explain all the observed variance in the target variable. While the inclusion of L1 and L2 regularization aids in mitigating overfitting and improves generalization, there are still complexities within the data that elude complete capture by the model. This shortfall suggests that there may exist latent factors influencing the target variable not accounted for in the feature set or not adequately modelled by the chosen architecture. Additionally, it underscores the inherent challenge in accurately modelling the intricacies of the underlying data generation process, prompting further exploration into potential model enhancements, feature engineering strategies, or data collection efforts to better elucidate and address the remaining variance.

## Implementation

Numerous studies have highlighted the variability in models utilized for genetic prediction, particularly concerning quantitative traits. It has become evident that a one-size-fits-all approach is not feasible in this context. Each trait presents unique characteristics and complexities, demanding tailored modelling strategies for optimal performance. Therefore, a systematic approach involving trial and error is imperative to identify the most suitable model for a specific dataset.

In the realm of genetic prediction, Single Nucleotide Polymorphisms (SNPs) serve as crucial markers. However, their efficacy is contingent upon their utilization within the appropriate population context. Understanding the population structure and genetic diversity is paramount in harnessing the predictive power of SNPs effectively. By accounting for population-specific factors, such as allele frequencies and linkage disequilibrium patterns, researchers can enhance the accuracy and reliability of genetic prediction models.

In summary, genetic prediction endeavours necessitate a nuanced and iterative approach, wherein diverse modelling techniques are explored, and SNPs are leveraged judiciously within the relevant population context. By embracing this iterative process and tailoring models to the intricacies of each trait and population, researchers can unlock deeper insights into genetic architecture and pave the way for more precise and personalized predictive analytics in genomics.

# Response to stakeholders

While Convolutional Neural Networks (CNNs) show promise as a tool for genetic prediction, further research is warranted to fully understand their capabilities and limitations in this domain. While CNNs have demonstrated effectiveness in various fields such as image recognition and natural language processing, their application to genetic prediction introduces unique challenges and considerations.

Firstly, the complex nature of genetic data, characterized by high-dimensional and heterogeneous features, necessitates careful exploration of CNN architectures and training strategies tailored to genetic prediction tasks. Factors such as the selection of input representations, model hyperparameters, and optimization techniques require thorough investigation to optimize CNN performance.

Additionally, the interpretability of CNN-based models in the context of genetic prediction remains an ongoing concern. Understanding how CNNs extract and prioritize features from genetic data is crucial for interpreting model predictions and identifying biologically relevant associations. Exploratory studies exploring the learned representations and decision-making processes of CNNs in genetic prediction tasks are essential for enhancing model interpretability and trustworthiness.

Moreover, the generalization and robustness of CNN models across diverse populations and genetic architectures require further examination. Evaluating the performance of CNNs on datasets representing different ancestries, genetic backgrounds, and trait complexities is essential for assessing their reliability and applicability in real-world scenarios.

Furthermore, ethical considerations, including privacy concerns and the responsible use of genetic information, must be carefully addressed in the development and deployment of CNN-based genetic prediction models. Ensuring transparency, fairness, and accountability in model development and implementation is paramount to building trust among stakeholders and safeguarding against potential harms.

While CNNs offer exciting opportunities for advancing genetic prediction research, their adoption in this domain necessitates rigorous investigation, validation, and consideration of ethical implications. Through collaborative interdisciplinary efforts and continued research, CNNs have the potential to revolutionize genetic prediction and contribute to personalized healthcare and precision medicine initiatives.

# End-to-end solution

After the birth of a cow, SNPs data can be systematically collected to characterize the genetic makeup of the animal. This genetic data, encompassing variations in the DNA sequence, serves as a valuable resource for predicting various phenotypic traits, including the potential milk yield of the cow.

By leveraging advanced machine learning models, such as predictive algorithms based on Convolutional Neural Networks (CNNs) or other regression techniques, the collected SNPs data can be inputted into the model to forecast the cow's future milk production capacity. These models analyze the genetic markers associated with milk production traits, identifying patterns and correlations that inform predictions about the cow's expected milk yield over its lifetime.

The ability to accurately predict the potential milk yield of a cow at an early stage offers numerous benefits for farmers and breeders. Firstly, it provides valuable insights into the genetic potential of individual animals, enabling informed decision-making regarding breeding strategies and management practices. Farmers can identify high-yielding cows early in their development, allowing for targeted investments in feed, housing, and healthcare to optimize their productivity.

Furthermore, predictive models empower breeders to make strategic decisions regarding the selection and mating of animals to improve overall herd performance. By prioritizing cows with the highest predicted milk production potential for breeding purposes, breeders can accelerate genetic progress and enhance the overall quality of the herd over successive generations.

Moreover, early predictions of milk yield enable farmers to implement proactive measures to address potential challenges or limitations in individual cow performance. For instance, cows identified as having lower-than-expected milk production may undergo tailored nutritional interventions or healthcare protocols to optimize their productivity and ensure their welfare.

Leveraging SNPs data and advanced predictive modelling techniques to forecast the milk yield of cows represents a powerful tool for enhancing agricultural productivity and sustainability. By harnessing the predictive power of genetics, farmers and breeders can make data-driven decisions that optimize herd management, improve breeding outcomes, and ultimately contribute to the efficiency and profitability of dairy operations.

# Bibliography

Dairy Australia. (n.d.). *AUSTRALIAN DAIRY COWS*. Retrieved from Dairy Australia: https://www.dairy.edu.au/information/australian-dairy-cows

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica, 23*(2), 143-149.

Pook, T., Freudenthal, J., Korte, A., & Simianer, H. (2020, November). Using Local Convolutional Neural Networks for Genomic Prediction. *Frontiers in Genetics, 11*.

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., . . . Li, J. (2015). Accuracy of Whole-Genome Prediction Using a Genetic Architecture-Enhanced Variance-Covariance Matrix. *G3: Genes, Genomes, Genetics*, 615.