Deep Learning-Based Prediction of Tomato Fruit Shapes from Genotype Data

Ramneek Kaur Arora
Computer Science
University of Western Ontario
London,Canada
rarora73@uwo.ca

***Abstract*—This directed study utilizes deep learning techniques for predicting tomato fruit shapes based on genotype data. By conducting an analysis of dataset from crosses between traditional and modern inbred varieties, we are utilizing it to run an established decoder architecture in replicating fruit shapes. Performance evaluation is being done by employing performance metrics to gauge their effectiveness in genetic data interpretation. The result reveals a notable variation in RMSE and MAE across varying training epochs, indicating the impact of training duration on decoder accuracy. This study highlights the vast potential of Deep Learning and Artificial Intelligence in the Agriculture setting and advocates for continued exploration in this field.**

***Index Terms*—Deep learning, genetic data, fruit shape prediction, machine learning.**

I. Introduction

Tomato fruit shapes are more than just pretty pictures; they include important information that is necessary to determine fruit quality and meet customer preferences. Understanding these forms and being able to forecast them with accuracy are critical skills in the complex field of agricultural genetics and breeding programs.

The study of plant morphology—including tomato fruit shapes—has been crucial for genetic research. Tomatoes are among the most studied plant species due to their enormous economic relevance and wide diversity of morphologies.

Traditionally, Tomato breeding strategies heavily relied on phenotype-based selection methods. Where breeders require years of expertise to carefully access fruit characteristics to guide their breeding decisions. However, with the rapid advancement of technology, particularly the emergence of artificial intelligence (AI) and deep learning, a new frontier emerges. Herein lies the potential to augment traditional breeding practices with computational methodologies, leveraging genotype data to precisely predict fruit shapes.

This study aimed at predicting tomato fruit shapes using deep learning techniques based on genotype data. By analyzing datasets from crosses between traditional and modern inbred varieties, the study employs established decoder architecture to replicate fruit shapes and assess their performance using various metrics. The goal is to demonstrate the feasibility and effectiveness of leveraging deep learning methodologies for fruit shape prediction tasks in agriculture. By training these models on large datasets of genotype-phenotype associations, the study aims to shed light on the genetic basis of fruit morphology and develop predictive tools to aid breeders in selecting plants with desired fruit shapes.

The report also discusses the tools and technologies utilized in the study, including deep learning frameworks like TensorFlow and Keras for model training and data analysis and manipulation libraries like NumPy and Pandas. Visualization tools such as Matplotlib and Seaborn are also employed for interpreting model outcomes and data patterns. By building upon existing literature and leveraging advancements in computational biology, the study aims to push the boundaries of predictive modeling in agriculture and contribute to the sustainability and efficiency of agricultural practices.

II. Background

A. Deep Learning for Shape Prediction from DNA

Deep learning for shape prediction from DNA leverages advanced computational models to interpret genetic information and predict the morphological traits of agricultural products, such as fruit shape in crops. By integrating genomic data with machine learning algorithms, researchers can generate detailed predictions about fruit appearance, enhancing breeding strategies and agricultural outcomes.

Deep Learning helps researchers with:

- Detailed Phenotypic Predictions: Deep learning models, particularly those involving decoders, are adept at translating complex genetic information (e.g., SNPs) into detailed phenotypic predictions like fruit shapes. These models can handle high-dimensional data and uncover patterns that are not readily apparent through traditional genetic analyses.

- Enhanced Breeding Efficiency: By accurately predicting fruit shapes from genetic data, deep learning enables breeders to select the most promising plants for further development early in the breeding cycle. This approach streamlines the selection process, reducing the time and resources spent on cultivating less desirable phenotypes.

- Innovations in Crop Development: Shape prediction drives innovations in crop development by facilitating the creation of new varieties with optimized shapes and other desirable traits. These advancements are vital for addressing global challenges like food security and adapting to changing climates. Additionally, optimized shapes can enhance marketability, offering economic benefits to farmers and stakeholders..

B. Tools and Technologies

The prediction of fruit shapes from DNA sequences leverages a suite of advanced computational tools and software packages, intricately designed to handle, analyze, and model the sophisticated interactions between genotypic and phenotypic data:

1) Deep Learning Frameworks: In the research, TensorFlow and Keras plays a vital role. TensorFlow offers a robust, scalable environment for training ML models, while Keras simplifies neural network construction with its intuitive API. We set the number of epochs and early stopping criteria to prevent overfitting. Callbacks, including early stopping, are employed to monitor training. Using the fit method, we train our model on dataset (dna train and imgs train), specifying parameters like batch size and validation data. This approach optimizes our model for genomic prediction tasks.

2) Data Analysis and Manipulation: We utilize Python libraries like NumPy and Pandas for efficient data analysis and manipulation. NumPy excels in numerical computing, enabling us to handle extensive genomic datasets seamlessly. Pandas provides versatile data structures and operations for manipulating numerical tables and time series, facilitating preprocessing and exploration of complex genomic datasets. To assess model performance, we use NumPy to compute RMSE and MAE scores by comparing predicted and actual images across the validation dataset. This systematic evaluation helps us refine and optimize our models effectively. Additionally, Pandas' data generation capabilities aid in creating random datasets for further experimentation.

3) Visualization Tools: In our research, we utilize Matplotlib and Seaborn for data visualization, aiding in understanding data patterns and model outcomes. These tools help create informative visualizations like line plots. Matplotlib plots RMSE and MAE scores against epochs, allowing us to track model performance evolution. Annotated points enhance plot interpretability. Together, Matplotlib and Seaborn provide robust visualization capabilities, enhancing data comprehension and facilitating communication of findings for research advancement.

III. Literature Review

This research builds upon the seminal work by Pérez-Enciso et al., who explored the potential of generating fruit shapes from DNA sequences [1]. Their innovative approach sets a foundation for my study, where we extend these methodologies to specifically predict tomato fruit shapes using deep learning techniques. Pérez-Enciso et al.'s work is critical as it demonstrates the feasibility of using genetic data to inform phenotypic outcomes in horticultural crops, an area that remains underexplored but holds significant promise for agricultural advancements.

Deep learning has increasingly become a powerful tool in genomic studies, providing the ability to handle complex and large datasets characteristic of genomic information. TensorFlow and Keras, tools mentioned by Abadi et al. and Chollet respectively, have been pivotal in developing computational models that efficiently process and learn from such data [2,3]. These frameworks are instrumental in my study, allowing for the creation and training of sophisticated neural network architectures capable of decoding genetic markers into phenotypic expressions like fruit shapes.

In the context of agriculture, the integration of genotypic and phenotypic data to predict plant characteristics offers transformative potential for breeding strategies. As demonstrated by Blanca et al., understanding genomic variation within crop species can lead to more targeted and efficient breeding programs [4]. My study leverages these insights by applying deep learning models to predict fruit morphology, thus enhancing the traditional breeding processes that have long relied on phenotype-based selection methods.

The use of image processing tools such as OpenCV, as mentioned by Bradski, facilitates the manipulation and analysis of plant images, which are critical for training deep learning models in my study [5]. These tools enable precise adjustments to the training images, ensuring that the models learn from clean and standardized data, which is vital for achieving high accuracy in phenotype prediction.

Furthermore, the predictive power of machine learning in agriculture is highlighted in the work by Cuevas et al., where they utilized deep learning to forecast plant traits in various environments [6]. This underscores the adaptive capability of these models to function under diverse agricultural settings, enhancing their practical application in real-world scenarios.

Lastly, I use a decoder model to forecast tomato fruit shape by integrating the methodological advances in computational biology, motivated by the generative models work of Goodfellow et al. Our method uses epochs, similar to the iterative adversarial training in GANs, where each dataset pass improves the model through several iterations[7], guaranteeing incremental learning and optimisation of the complex genotype-phenotype correlations , our structured training aims to capture the genetic patterns determining tomato fruit shape, so providing the generation of realistic and accurate phenotypic predictions.

IV. Methodology

A. Data Collection and Preparation

Dataset Composition: The dataset comprises a diverse collection of 353 tomato images obtained from 129 experimental crosses involving both traditional and modern inbred lines. These images serve as the primary source of phenotypic data, capturing the variability in tomato fruit shapes across different genetic backgrounds.

In conjunction with the image data, genetic information is obtained through genotyping by sequencing (GBS), yielding a set of 68 segregating single nucleotide polymorphisms (SNPs). These SNPs are selected based on their association with fruit shape-related candidate genes, providing a genetic foundation for the predictive modeling process.

Additionally, the dataset incorporates supplementary phenotypic data derived from biochemical, color, and morphological metrics of hybrid tomato fruits. These metrics, totaling 48 in number, are measured carefully and recorded for a subset of 32 founder lines. Through linear regression, these metrics are extrapolated for the entire dataset, enriching the phenotypic information available for modeling purposes.

a) Image Processing: Pre-processing of tomato images is carried out to standardize their attributes. This involves resizing the images to a uniform size, adjusting color balance to mitigate variations in lighting conditions, and cropping to remove extraneous background noise. By standardizing the images, the dataset achieves homogeneity, enhancing the model's ability to learn and generalize across diverse fruit shapes.

B. Neural Network Architecture

Decoder Architecture:

- **Input Layer**: The input layer of the decoder receives a vector representing the SNP data, encapsulating the genetic makeup influencing fruit shapes. Each element of this vector corresponds to a specific SNP, capturing the variations in genetic sequences across different tomato varieties.
- **Dense Layer**: Following the input layer, a dense layer with a high number of neurons processes the genetic data. This layer acts as a feature extractor, transforming the genetic information into a rich and abstract feature set that captures the complex interactions between genetic markers and fruit morphology.
- **Reshape Layer**: The output from the dense layer is reshaped to match the dimensions required for subsequent convolutional operations. This reshaping step prepares the data for the convolutional transpose layers, facilitating the reconstruction of the image.
- **Convolutional Transpose Layers**: The core of the decoder architecture comprises convolutional transpose layers. These layers play a crucial role in reconstructing the image from the encoded genetic data. By progressively upsampling and convolving the data, these layers generate finer details, ultimately producing an output image that closely resembles the original tomato fruit shape.

Training the Model:

- **Loss Function:** During training, the model minimizes the mean squared error (MSE) between the predicted and actual images. This loss function quantifies the discrepancy between the reconstructed fruit shape and the ground truth, guiding the optimization process through backpropagation.

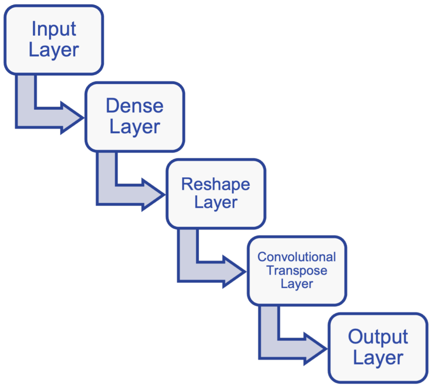
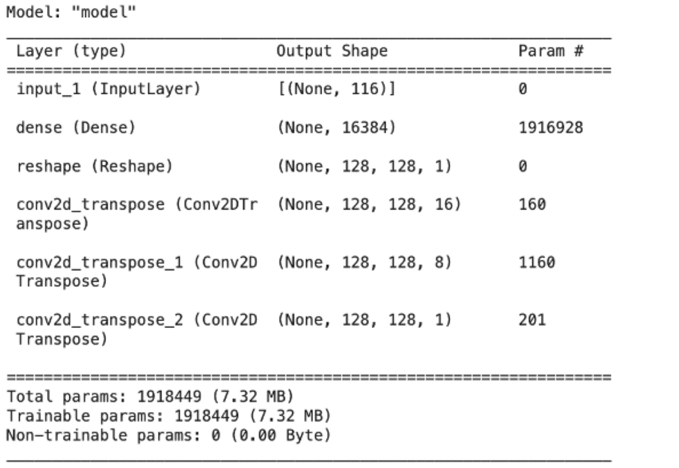


Fig. 1. Basic Idea of the Decoder



```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 116)]             0

 dense (Dense)               (None, 16384)             1916928

 reshape (Reshape)           (None, 128, 128, 1)       0

 conv2d_transpose (Conv2DTr  (None, 128, 128, 16)      160
 anspose)

 conv2d_transpose_1 (Conv2D  (None, 128, 128, 8)       1160
 Transpose)

 conv2d_transpose_2 (Conv2D  (None, 128, 128, 1)       201
 Transpose)

=================================================================
Total params: 1918449 (7.32 MB)
Trainable params: 1918449 (7.32 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 2. Decoder Architecture

- **Optimizer**: To optimize the model's parameters, adaptive optimizers such as Adam or RMSprop are employed. These optimizers dynamically adjust the learning rate based on the gradients of the loss function, enabling efficient convergence and mitigating issues associated with sparse gradients.

C. Model Training and Validation

Training Setup: This phase involves training the model at multiple epochs(20,30,40,50, and 60), during which the model iteratively learns to minimize the loss function. The partition of the dataset is done in a way to allocate 80% of the data for training the model and rest 20% is set aside for validation. This partition was done previously and I am using it as is. In addition, I experimented with a different partitioning strategy where I utilized 70% of data for training instead of 80%, and rest 30% for validating. The finding indicates that the change in partitioning size influences the performance metrics where we see a rise in RMSE and MAE score. Which means the original splitting strategy yielded the better performance of the model and hence we used it in the further research.

Performance Evaluation:

- **Quantitative Metrics**: The performance of the model is evaluated using quantitative metrics such as root mean squared error (RMSE) and mean absolute error (MAE). These metrics provide numerical insights into the accuracy and precision of the model's predictions, quantifying the discrepancies between the predicted and actual fruit shapes.
- **Qualitative Assessment**: Additionally, qualitative assessment through visual comparison of predicted and actual tomato fruit images is conducted. This visual inspection allows researchers to assess the realism and fidelity of the generated fruit shapes, providing complementary insights to the quantitative metrics.

D. Random Data Generation:

Continued the directed study, I tried training the model with random input data to check how the decoder performs on randomly generated data.

The random data generation process involves several key functions tailored to produce diverse datasets, encompassing both structured and unstructured data. Firstly, the ‘generate_ped_2‘ function constructs pedigree data, describing individuals within a population. This function generates a pandas DataFrame with attributes such as gender and genetic markers, crucial for characterizing each individual's genetic profile. Complementing this, the ‘generate_qtn_2‘ function synthesizes genotypic data, furnishing insights into genetic traits across the population. These functions collectively form the backbone of our genetic dataset, providing a foundation for multi-modal analysis.

Incorporating structural context, the ‘generate_contours‘ function fabricates contour data, representing geographical features pertinent to the study. This data adds a spatial dimension to our dataset, facilitating the integration of genetic information with environmental factors. Additionally, the ‘generate_random_array‘ function contributes by generating random coordinate arrays, further diversifying the spatial data landscape. By combining these functions, we construct a comprehensive dataset capturing the complexity of genetic and environmental interactions.

The integration of image data is critical in enriching our dataset with visual information. Leveraging contour data, the ‘Imgset_2‘ generation process crops images corresponding to geographical contours associated with each individual. This fusion of genetic and visual data empowers our model with a holistic understanding of the underlying phenomena. Through the arrangement of these functions, we curate a robust dataset primed for model training and evaluation.

V. RESULTS

The study deployed a deep-learning decoder to predict the shapes of tomato fruits using genotypic data marked by SNPs. We evaluated the model's accuracy through statistical and visual methods. Quantitative metrics such as RMSE and MAE gauged precision, while visual assessments allowed us to compare predicted shapes against actual images to confirm the model's effectiveness. This dual approach validated the model's capability to both learn detailed shapes from the data and generalize well to new, unseen genotypes.

1) Quantitative Metrics: The decoder was trained over varying epochs(20 to 60) to determine the optimal training duration for maximum accuracy. Results indicated a significant impact of training duration on model performance.

- **Root Mean Squared Error (RMSE)**: In the study, the Root Mean Squared Error (RMSE) served as a key metric to evaluate the accuracy of our deep learning model in predicting tomato fruit shapes from genetic data. At the initial 20 epochs, the RMSE was recorded at 32.97 for images sized at 128x128 pixels, which indicated that while the model was starting to learn, there was still substantial room for improvement. As the training progressed to 40 epochs, there was a slight decrease in RMSE to 32.55, demonstrating a positive trend with prolonged training and suggesting that the model was gradually refining its predictive accuracy. Most notably, upon extending the training to 60 epochs, the RMSE showed an improvement, reducing to 31.78.[Fig 3.] This marked decrease highlights the benefits of extended training sessions, clearly showing that longer training periods can greatly enhance the model's performance in accurately predicting fruit shapes

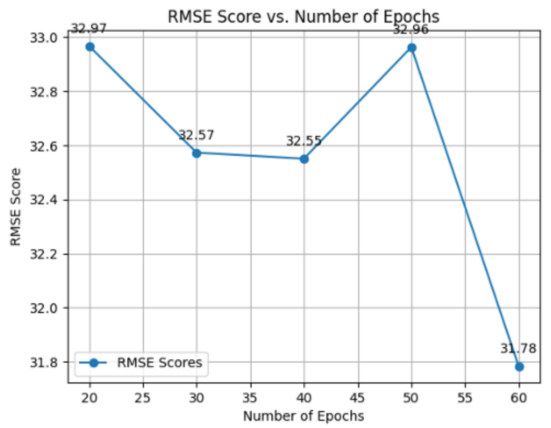


Fig. 3. RMSE Score from Epochs 20 to 60

- **Mean Absolute Error (MAE)**: The Mean Absolute Error (MAE) scores obtained during the model's training phase provide insightful revelations into the model's learning trajectory over a range of epochs[Fig 4.]. The initial MAE score at 20 epochs stood at 10.28, indicative of the model's nascent stage in learning from the data. Progressing through the epochs, the model experienced a slight fluctuation in performance, with a peak MAE score of 10.33 at 40 epochs, suggesting variability in learning. However, upon extending the training duration, a significant enhancement in performance was observed. Notably, at 60 epochs, the model achieved a MAE score of 9.22, the lowest across the training spectrum, reflecting a robust improvement in the model's predictive accuracy. The decline in the MAE score from the earlier epochs to 60 epochs underscores the critical importance of prolonged training in the model's ability to predict tomato fruit shapes more precisely from the genetic data.

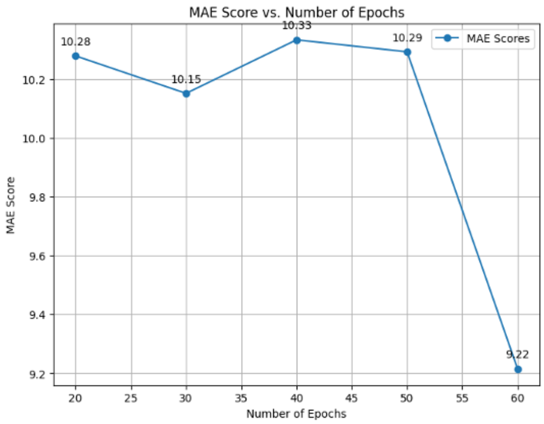


Fig. 4. MAE Score from Epochs 20 to 60

- **Effect of Image Size on Model Performance:** In my experiment, I explored the impact of different image sizes on the performance of my model, focusing on resolutions of 128x128 and 256x256 pixels. Using RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) as evaluation metrics, I found that the model exhibited better performance with the smaller image size. Across various epochs, the RMSE values consistently rose for the 256x256 pixel images, indicating challenges in handling higher resolutions. Similarly, the MAE scores for the larger images remained consistently higher compared to the smaller ones. These results suggest that the model struggles to effectively learn from higher resolution data, highlighting potential limitations in its architecture or the need for further optimization techniques to address the complexities introduced by larger images.

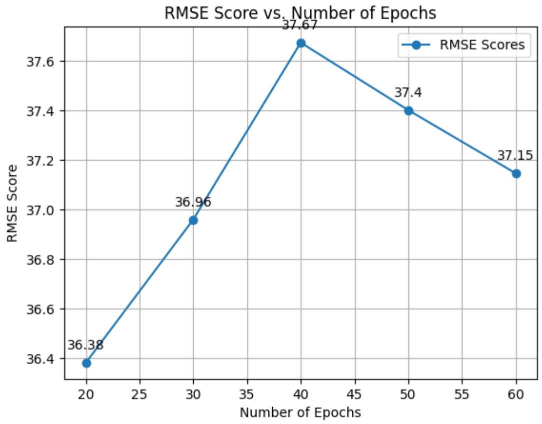


Fig. 5. RMSE Score for Epochs 20 to 60 (256x256)

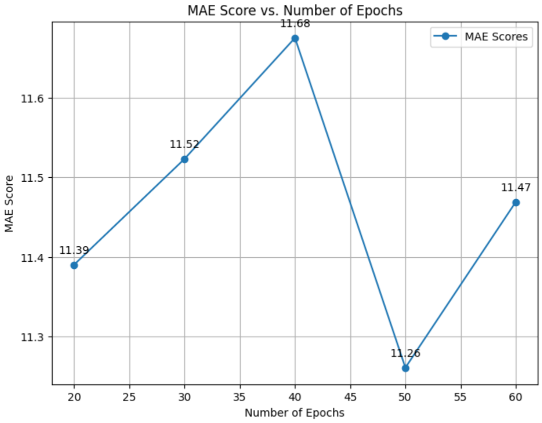


Fig. 6. MAE Score for Epochs 20 to 60 (256x256)

2) Visual Assessment: Visual comparisons of predicted images against actual images were conducted for models trained at different epochs, revealing insights into model performance.

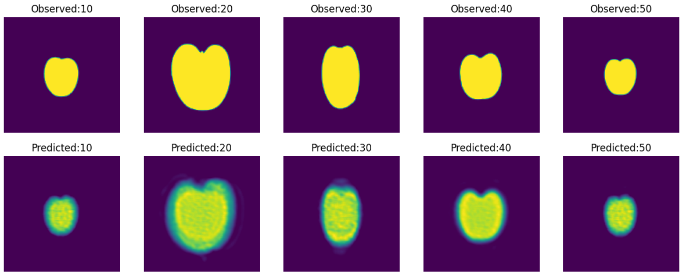


Fig. 7. Tomato fruit shape at Epochs 20

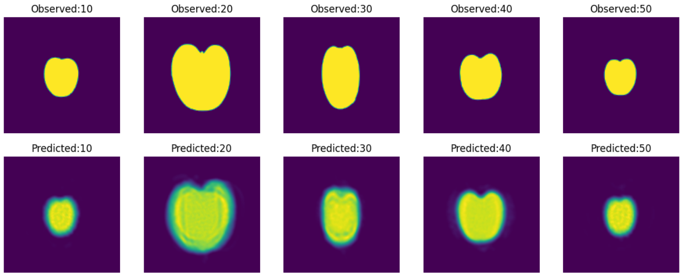


Fig. 8. Tomato fruit shape at Epochs 60

Figures 7 and 8, show the tomato fruit shape at two different epochs, 20 and 60. We can see a slight improvement in the generated image at 60 epochs. The model trained at 20 epochs produced predictions that, while capturing the general shape of the fruit, lacked definition and depth, presenting blurred edges and an overall lack of detail. This shows the model's early stage of learning, where it began to understand the basic structure but was not yet adept at rendering the finer aspects[Fig.7]. The predictions underwent a little transformation after increasing the training to 60 epochs. The contours of the predicted fruit shapes become slightly sharper, and the color fidelity also improved, closely mirroring the clarity observed in the actual fruit images[Fig.8]. The results demonstrate the critical role that extended training plays in deep learning models tasked with complex image predictions. As the number of epochs increased the model's ability to process and replicate the characteristics of the input image also increased. This progression highlights the importance of ample training time for achieving better results in predictive modeling.

3) Result of Random Data Generation:: We ran into a technical issue while creating plots for predicted shapes based on random data. An attempt was made to access an array's out-of-bounds index, which resulted in an IndexError. The mistake occurred specifically when the code tried to access index 10 of an array with a size of only 3. This problem suggests that our charting routine's indexing logic or input data were not in line with the size and actual data structure. This ultimately led to the plots that were meant to be produced not developed. This inaccuracy highlights the need for thorough evaluation of automated data processing processes, data handling and indexing systems in order to avoid reoccurring problems in subsequent studies.

4) Model Optimization Insights:

– The study underscored the significance of longer training periods in improving model performance, particularly evident in the enhanced detail and accuracy of predicted fruit shapes.

– Handling larger image resolutions effectively may necessitate adjustments in model architecture or training strategies, such as increased depth or augmented datasets.

VI. CONCLUSION

In this study, we explored the use of deep learning models to predict tomato fruit shapes from genotype data. The advanced neural network architectures employed demonstrated significant potential in decoding genetic information into phenotypic expressions, underscoring the transformative power of machine learning in agricultural genetics.

Our findings revealed that while the model excelled when working with familiar, well-characterized training data, it faced significant challenges with randomly generated inputs. This issue highlights a critical limitation: the model's inability to effectively handle data that deviates from realistic genetic patterns.

To enhance the model's robustness and its ability to handle anomalous data, we propose refining the data augmentation techniques used during training. Rather than relying on completely random genetic inputs, introducing controlled, realistic variations into the existing genetic sequences may prove more effective. This method would involve simulating mutations and other genetic variations that could realistically occur, allowing us to test how well the model adapts to these changes. Such a strategy would help improve the model's generalizability across different genetic scenarios and its performance in real-world applications.

Using genetic simulation tools to create a wide range of reasonable genetic scenarios for testing would provide a more thorough assessment of the model's performance under varied conditions. This approach not only tests the resilience of the model but also ensures that it can operate effectively across a broader spectrum of genetic diversity, which is critical in the rapidly evolving field of agricultural genetics.

In conclusion, our study affirms the potential of deep learning for enhancing phenotypic prediction in agriculture but also emphasizes the need for continued research and development. By further refining model training methodologies and improving data handling techniques, we can significantly boost the predictive capabilities of artificial intelligence in breeding programs. Such advancements will not only enhance crop yield and quality but also contribute to the sustainability and efficiency of agricultural practices in a changing global environment.

REFERENCES

[1] M. Pérez-Enciso, C. Pons, A. Granell, S. Soler, B. Picó, A.J. Monforte, L.M. Zingaretti, "Computer generation of fruit shapes from DNA sequence," 2022.

[2] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems." Available online: https://www.tensorflow.org/overview.

[3] F. Chollet, "Deep Learning with Python," Manning Publications, 2021.

[4] J. Blanca et al., "Genomic variation in tomato, from wild ancestors to contemporary breeding accessions," *BMC Genomics*, vol. 16, 2015. doi: 10.1186/s12864-015-1444-1.

[5] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[6] J. Cuevas et al., "Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials," *G3: Genes—Genomes—Genetics*, vol. 9, no. 9, 2019. doi: 10.1534/g3.119.400493.

[7] I. Goodfellow et al., "Generative Adversarial Nets," ArXiv, 2014. Available online: https://arxiv.org/abs/1406.2661.