Results of Design of Experiments (DOE) and Post-DOE Studies

Hossein Beidaghydizaji

Abstract

The DOE and Post-DOE investigations aim to systematically optimize CNN hyperparameters to improve classification performance between male and female facial images employing a small image dataset. This project uses free portrait images downloaded from Pixabay, which are licensed for free non-commercial use under the Pixabay License. The collected and used local dataset includes 471 images for the female class and 472 images for the male class. As the employed dataset is small, data augmentation is also considered before creating the CNN model to enhance the model accuracy. Further information regarding the data augmentation is available in the uploaded notebook files of the selected design points or the results presented in PDF format in the result section in GitHub repository of the project. The DOE approach chosen is a Central Composite Design (CCD), suitable for capturing both linear and nonlinear (quadratic) relationships. Two core responses were studied:

- Model complexity via trainable parameters
- Model performance via F1-score accuracy

The study provides quantitative insight into how hyperparameters such as learning rate, dropout, number of Conv2D layers, and filter sizes impact model behavior and predictive performance.

The process of DOE model design as well as the project introductory information can be found in the PDF file named "Project Information & Introduction" in the project repository in the GitHub.

1. Results and Discussion

1.1. Exploratory Data Analysis (EDA)

The heatmap of the correlation between DOE hyperparameters and the two main responses, i.e., trainable parameters and F1 score accuracy is presented in Figure 1. There is a strong positive correlation between architectural parameters, such as the number of Conv2D layers or filters, and the trainable parameter count, indicating that deeper and wider networks, respectively, inherently produce more model complexity. The learning rate however is more negatively correlated with F1 score accuracy, highlighting that the learning rate's influence is more crucial for both model convergence and generalization. There was little relationship between the dropout and the explanatory levels, indicating little effect.

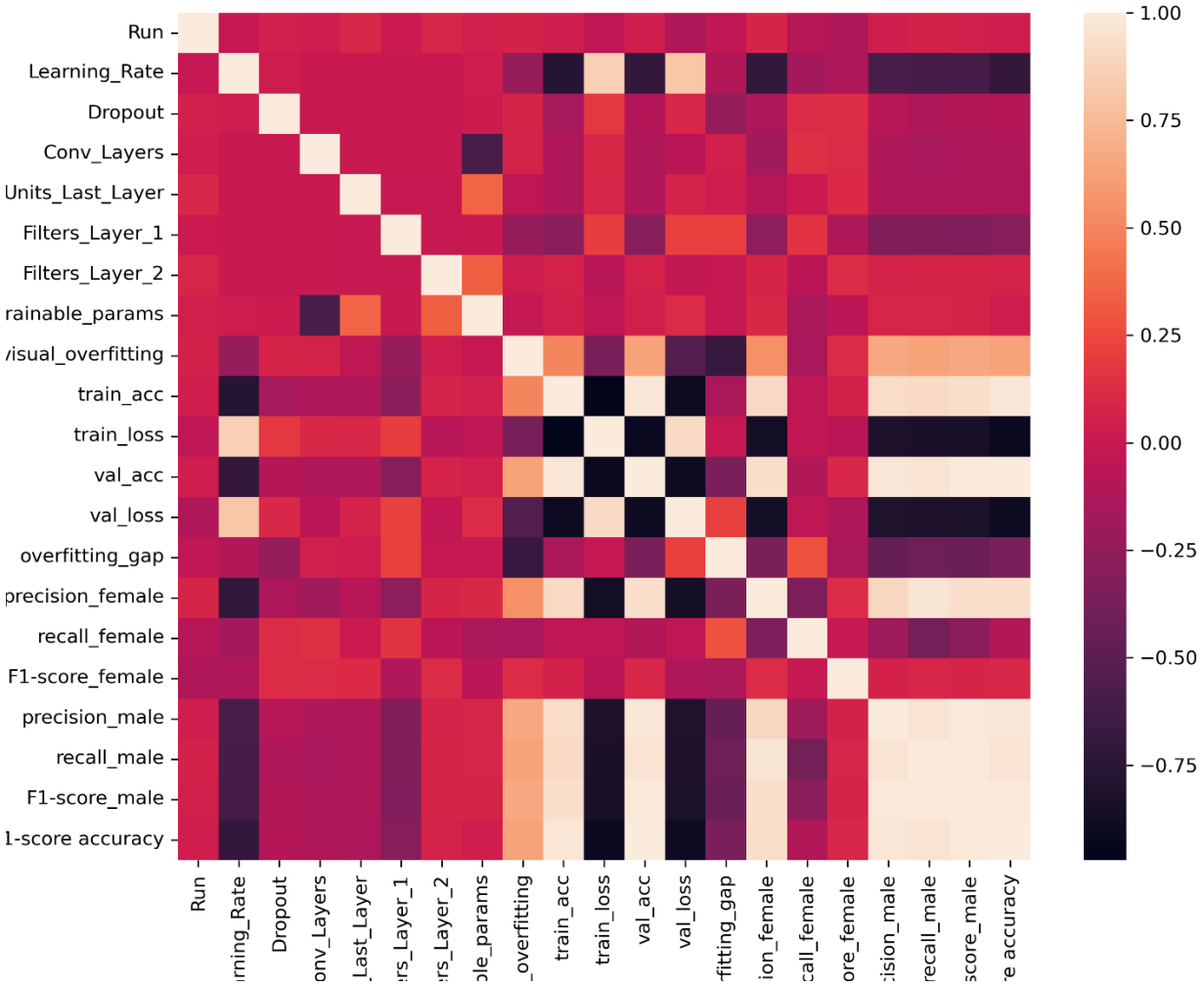


Figure 1. Heatmap of DOE factors and responses

As shown in Figure 2, the scatter plot of learning rate versus F1 score accuracy reveals a clear non-linear relationship. The highest F1 results are observed at moderate learning rates between 0.0005 and 0.005. Lower rates, such as 0.00025, have been demonstrated to lead to underfitting that is characterized by slow and inefficient convergence. In contrast, higher rates approaching 0.01 have been observed to instabilize training and result in divergence. This underlines the importance of carefully calibrating the learning rate to obtain optimal model performance.

According to the results, the maximum achieved F1-scores accuracy among all 81 DOE runs relates to the run number 6 which is 87.00%, and average F1-score = 76.38%. Detailed numerical values can be found in the attached Excell sheet of the recorded results. Figure 3 demonstrates the relationship between the model complexity factors (e.g. number of filters in second Conv2D layer, number of Conv2D layers, and number of units before the last dense layer) and the trainable parameters. Generally, an increase in trainable parameters is observed as the model complexity rises, although without detailed DOE analysis including the interaction effects of the factors, it is difficult to conclude from these simple results. It should also be considered that increasing the model complexity imposes a substantial computational burden and the risk of overfitting if it is not properly managed.

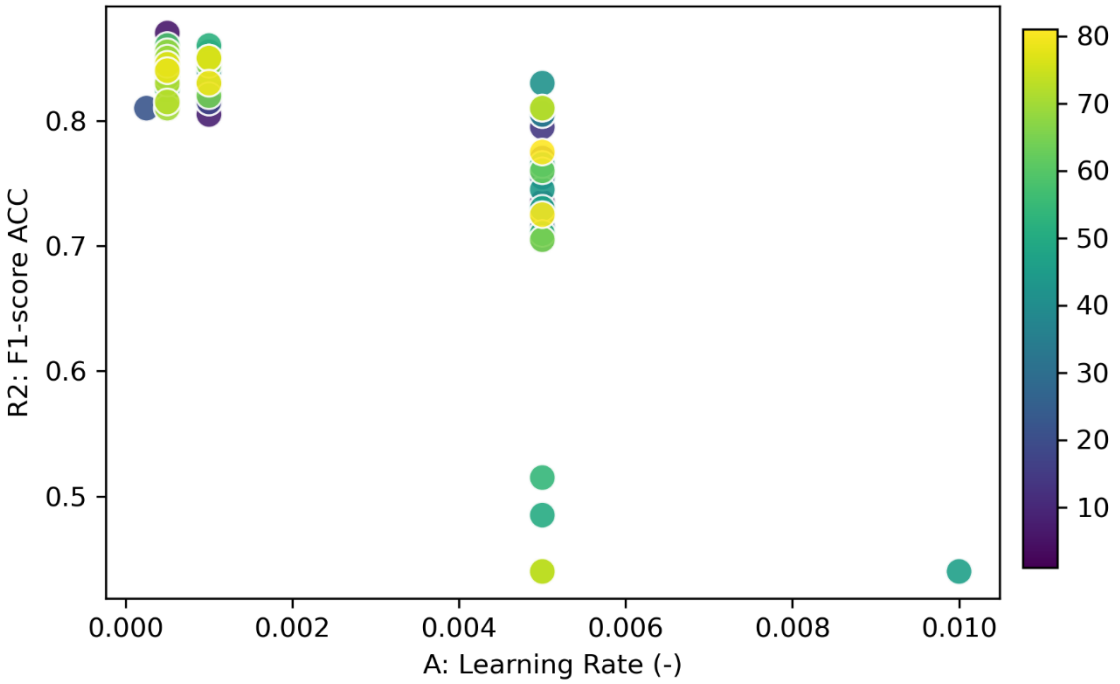


Figure 2. Scatterplot of learning rate vs. F1-score accuracy

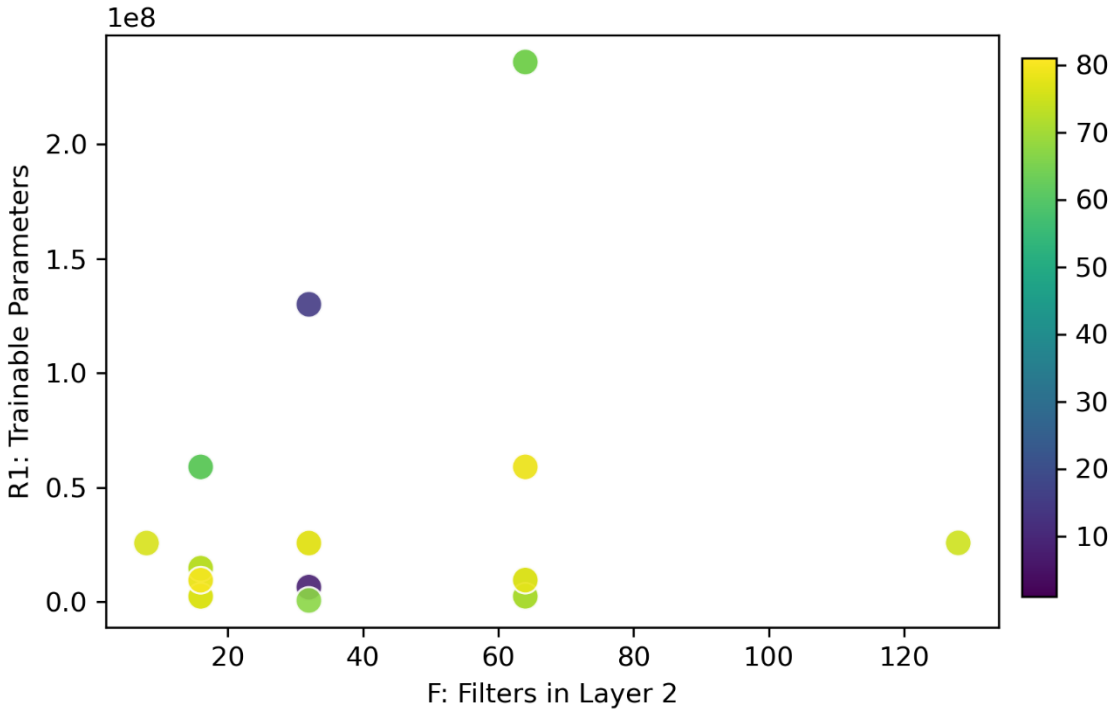


Figure 3a. Scatterplot of Number of Filters in Layer 2 vs. trainable parameters

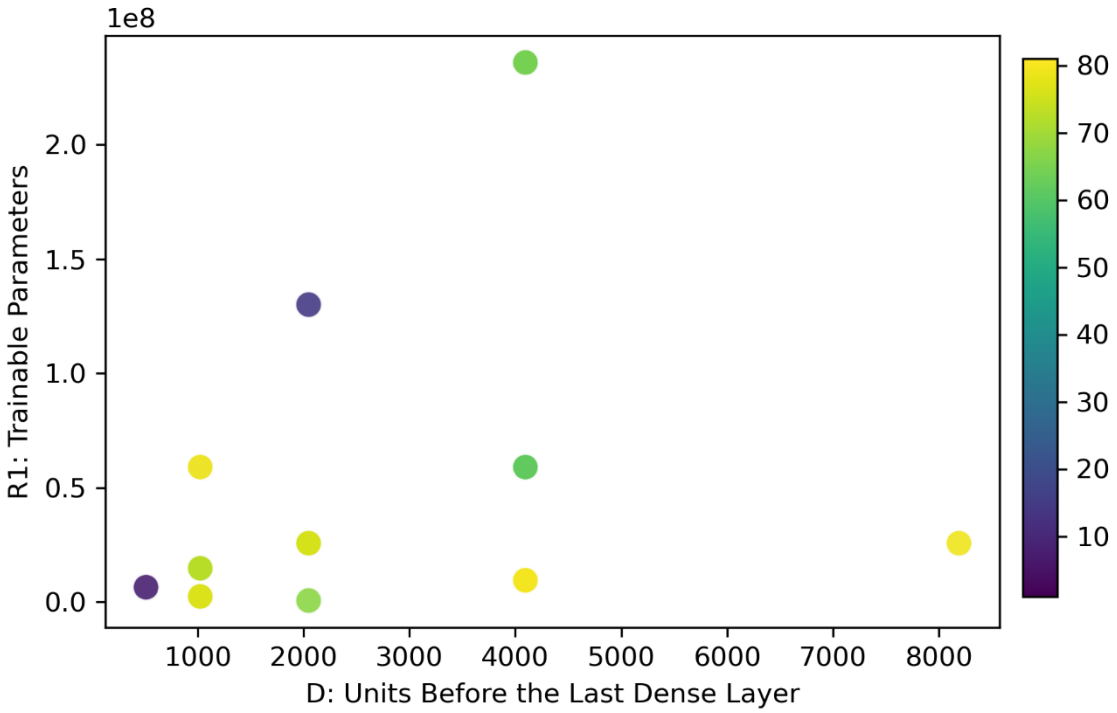


Figure 32. Scatterplot of Number of Units Before the Last Layer vs. trainable parameters

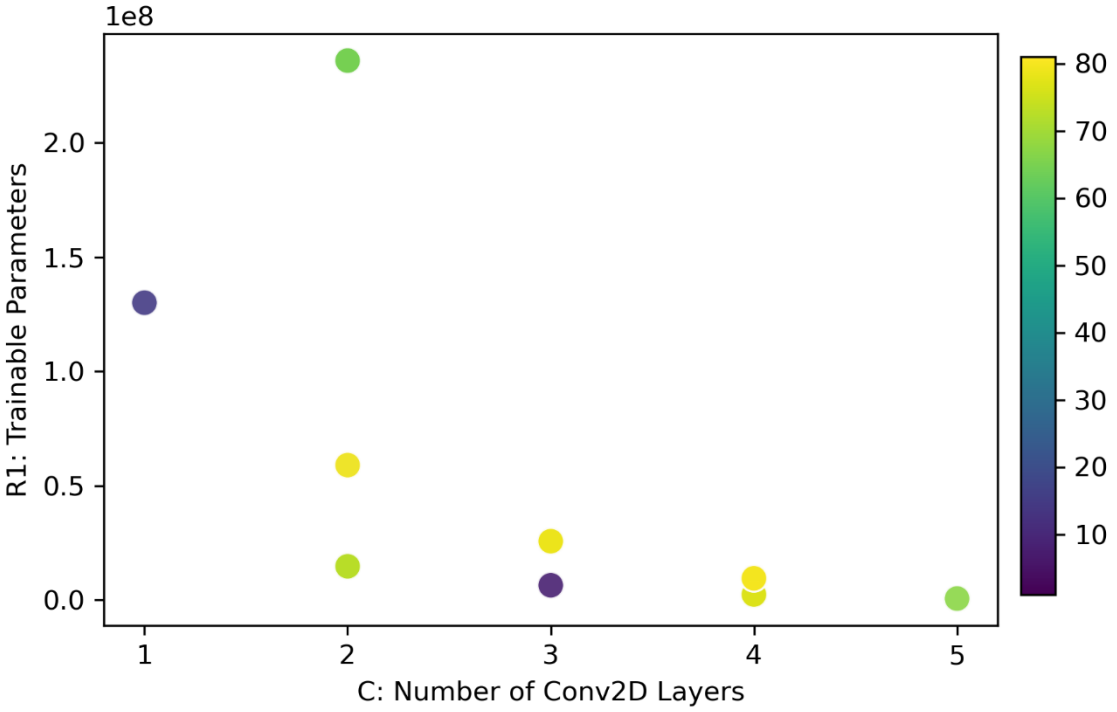


Figure 3c. Scatterplot of Number of Con2D Layers vs. trainable parameters

Figure 4 illustrates a perfect correlation between validation accuracy and F1-score accuracy. A strong positive relationship is evident, indicating that validation accuracy is a reliable early indicator of final model performance. This supports the strategy of employing early stopping based on validation accuracy, as implemented in the post-DOE optimization phase. Additional scatterplots can be found later in the appendix section.

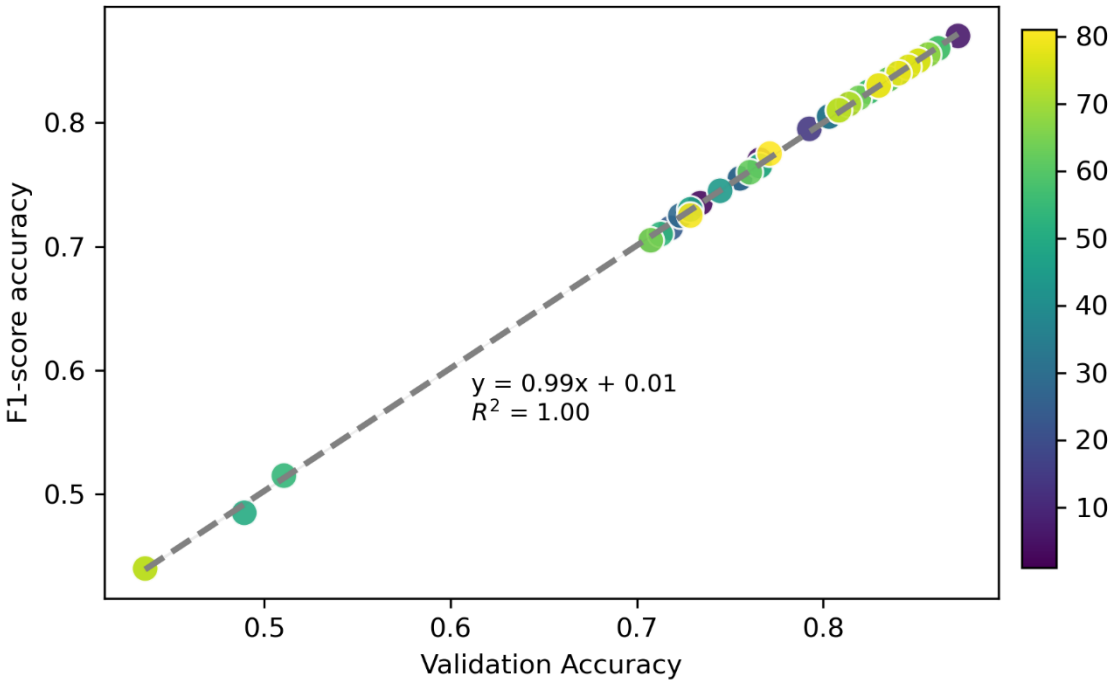


Figure 4. Scatterplot of validation vs. F1-score accuracy

More details regarding the EDA are available in the attached EDA notebook in the result section of the project GitHub repository.

1.2. DOE Analysis of First Response: Trainable Parameters

Model selection based on sequential p-values and model adequacy metrics (Adjusted R^2 and Predicted R^2), presented in Table 1, indicated that both 2FI and quadratic models are suitable, with 2FI offering slightly better predictive reliability. ANOVA results, which are reported in Table 2, reveal that the number of Conv2D layers (C), units before the last layer (D), and filters in the second layer (F) have significant main effects ($p < 0.0001$). Significant interactions (CD, CF, DF) and quadratic terms (C^2, D^2, F^2) were also identified, confirming nonlinear complexity growth patterns.

The final regression equation for trainable parameters is:

Trainable Params = +1.215E+08 − 7.475E+07C + 2.859E+07D + 2.835E+07F − 1.601E+07CD − 1.712E+07CF + 1.323E+07DF + 5.234E+06C^2 − 5.587E+06D^2 − 4.966E+06*F^2

The results emphasize that Conv2D layers have the strongest individual effect on complexity, with interaction and quadratic terms significantly contributing to the model.

Table 1. Model Selection for First Response

Source	Sequential p-value	Adjusted R^2	Predicted R^2	Info
Linear	< 0.0001	0.5469	0.4774	
2FI	**< 0.0001**	**0.8733**	**0.8110**	**Suggested**
Quadratic	< 0.0001	0.9202	0.7263	Suggested
Cubic		1.0000		Aliased

Table 2. ANOVA Summary for First Response

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	3.476E+17	9	3.862E+16	140.09	< 0.0001	**significant**
C-Conv Layers	1.924E+17	1	1.924E+17	697.91	< 0.0001	
D-Units Last Layer	1.605E+16	1	1.605E+16	58.22	< 0.0001	
F-Filters Layer 2	1.578E+16	1	1.578E+16	57.23	< 0.0001	
CD	4.253E+16	1	4.253E+16	154.25	< 0.0001	
CF	4.858E+16	1	4.858E+16	176.20	< 0.0001	
DF	1.848E+16	1	1.848E+16	67.02	< 0.0001	
C^2	3.426E+15	1	3.426E+15	12.43	0.0007	
D^2	6.769E+15	1	6.769E+15	24.55	< 0.0001	
F^2	5.349E+15	1	5.349E+15	19.40	< 0.0001	
Residual	1.958E+16	71	2.757E+14			
Lack of Fit	1.958E+16	67	2.922E+14			**insignificant**
Pure Error	0.0000	4	0.0000			
Cor Total	3.672E+17	80				

Figure 5 shows the normal probability plot of residuals for trainable parameters. The residuals largely align along the reference line, supporting the normality assumption necessary for valid statistical inference. Figure 6 presents the residuals versus run order plot. Residuals are randomly dispersed without any evident patterns, confirming the absence of autocorrelation across the run sequence and reinforcing the successful randomization during experimentation.

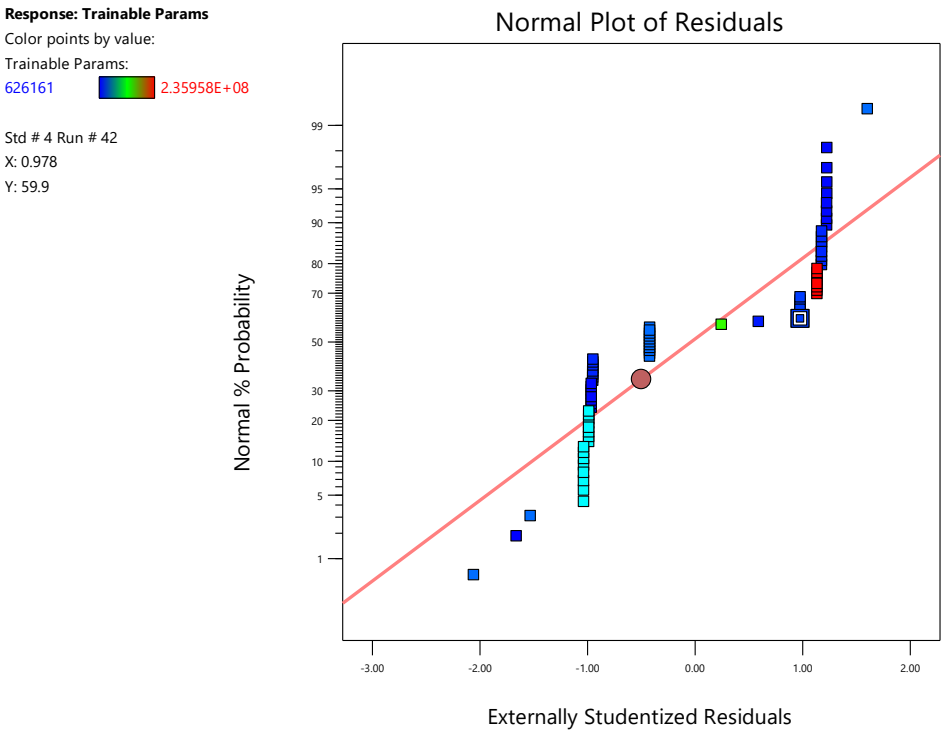


Figure 5. Normal Probability Plot of Residuals for Trainable Parameters

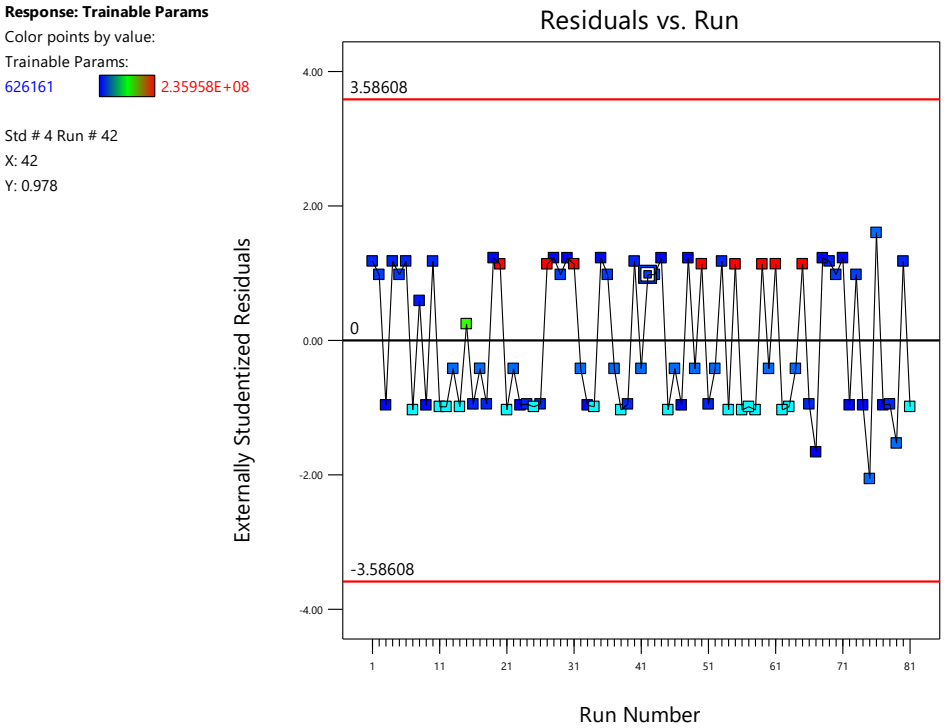


Figure 6. Plot of Residuals vs. Run Order for Trainable Parameters

Following the validation of model assumptions, the effects of key factors and their interactions on the complexity of the CNN (in terms of trainable parameters) were explored using contour plots and 3D surface plots:

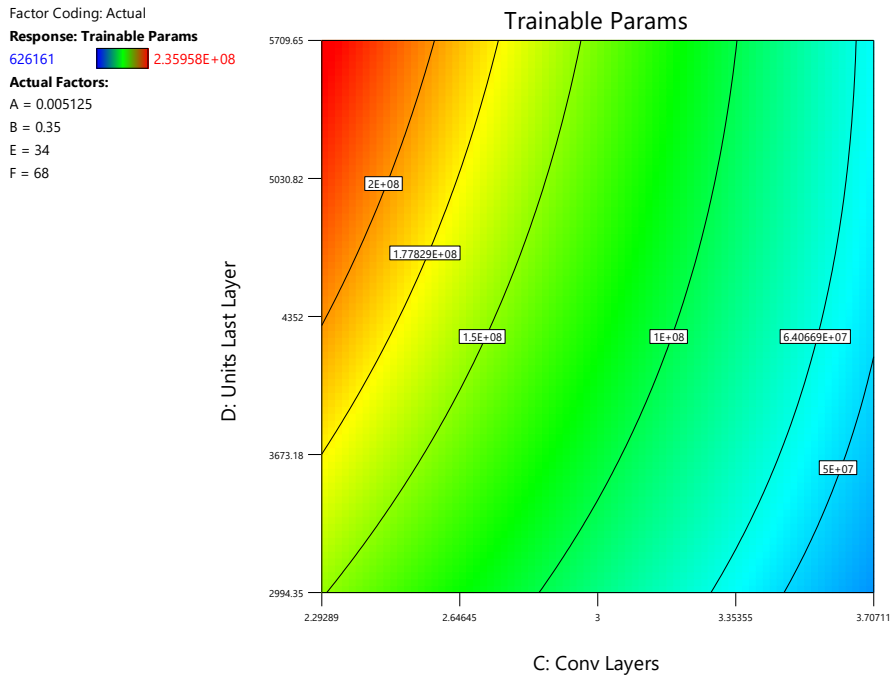


Figure 7. Contour Plot of Trainable Parameters vs. Conv2D Layers and Units Before Last Layer

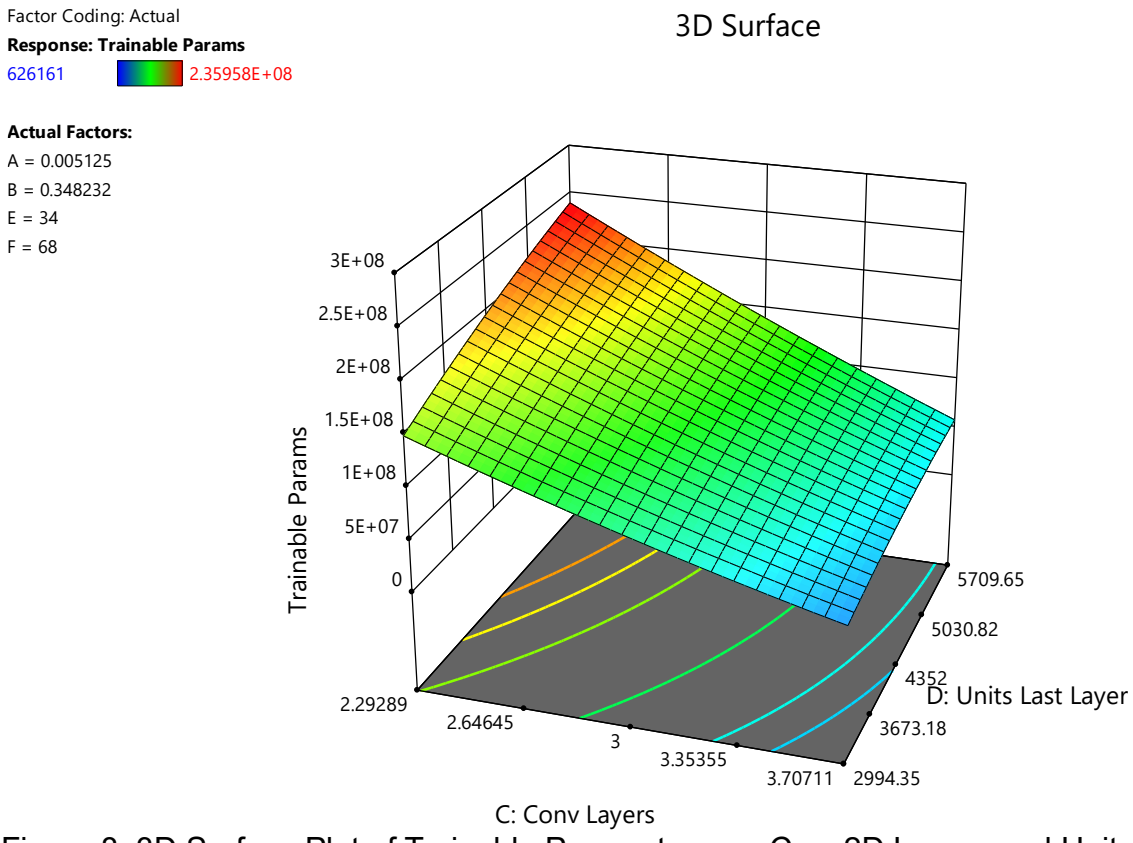


Figure 8. 3D Surface Plot of Trainable Parameters vs. Conv2D Layers and Units Before Last Layer

Figure 7 and Figure 8 illustrate the interaction between Conv2D layers and units before the last layer. Both plots indicate a strong synergistic effect: as both the number of convolutional layers and the units before the final dense layer increase, model complexity escalates significantly. This is consistent with the significant CD interaction term identified in the ANOVA (Table 2).

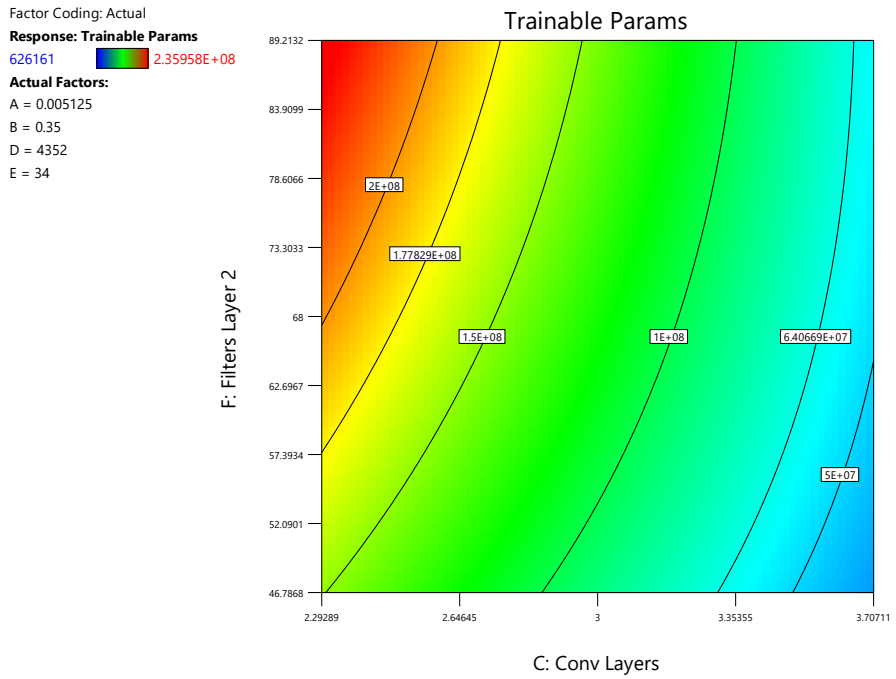


Figure 9: Contour Plot of Trainable Parameters vs. Conv2D Layers and Filters in Second Layer

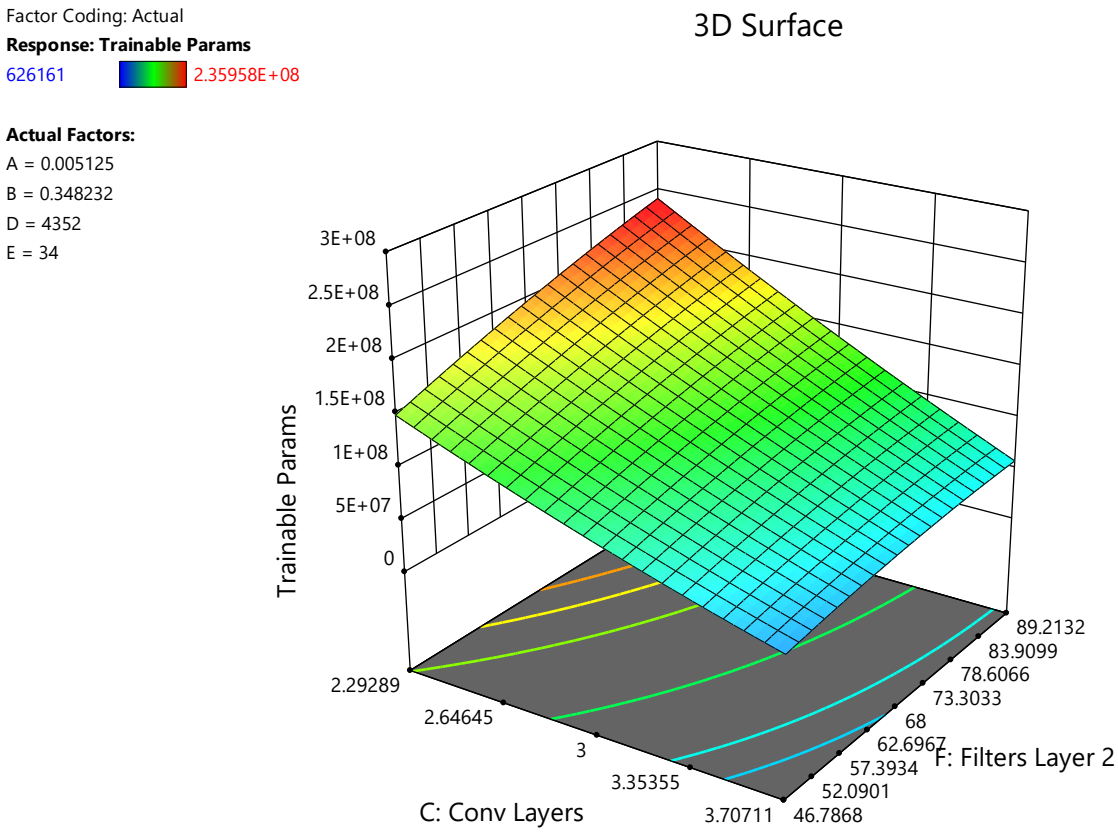


Figure 10. 3D Surface Plot of Trainable Parameters vs. Conv2D Layers and Filters in Second Layer

Figure 9 and Figure 10 examine the relationship between Conv2D layers and filters in the second layer. The contour and surface plots again reveal a pronounced interaction, with trainable parameters increasing rapidly when both architectural depth and width grow, affirming the statistical significance of the CF interaction.

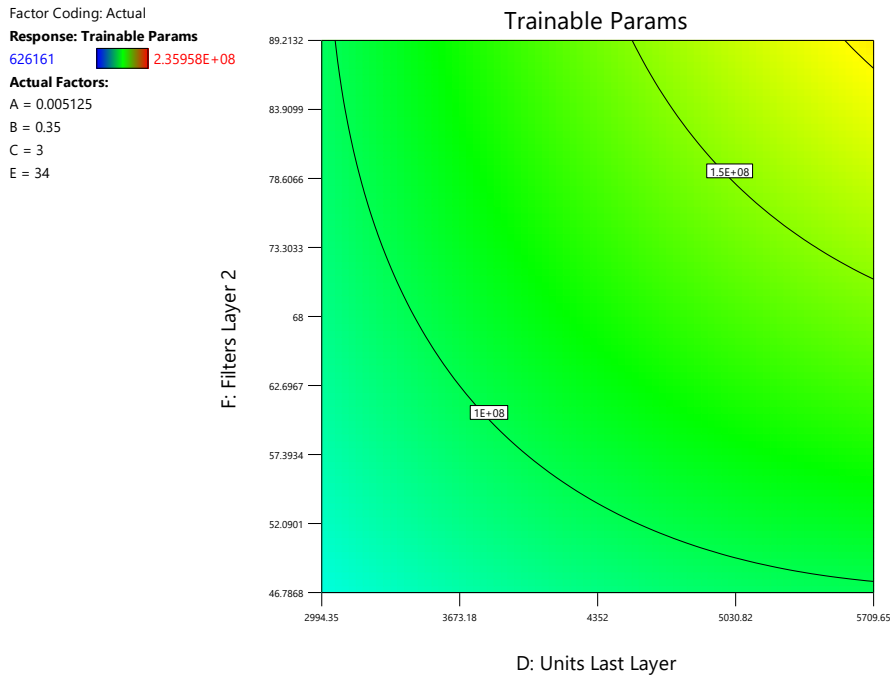


Figure 11: Contour Plot of Trainable Parameters vs. Units Before Last Layer and Filters in Second Layer

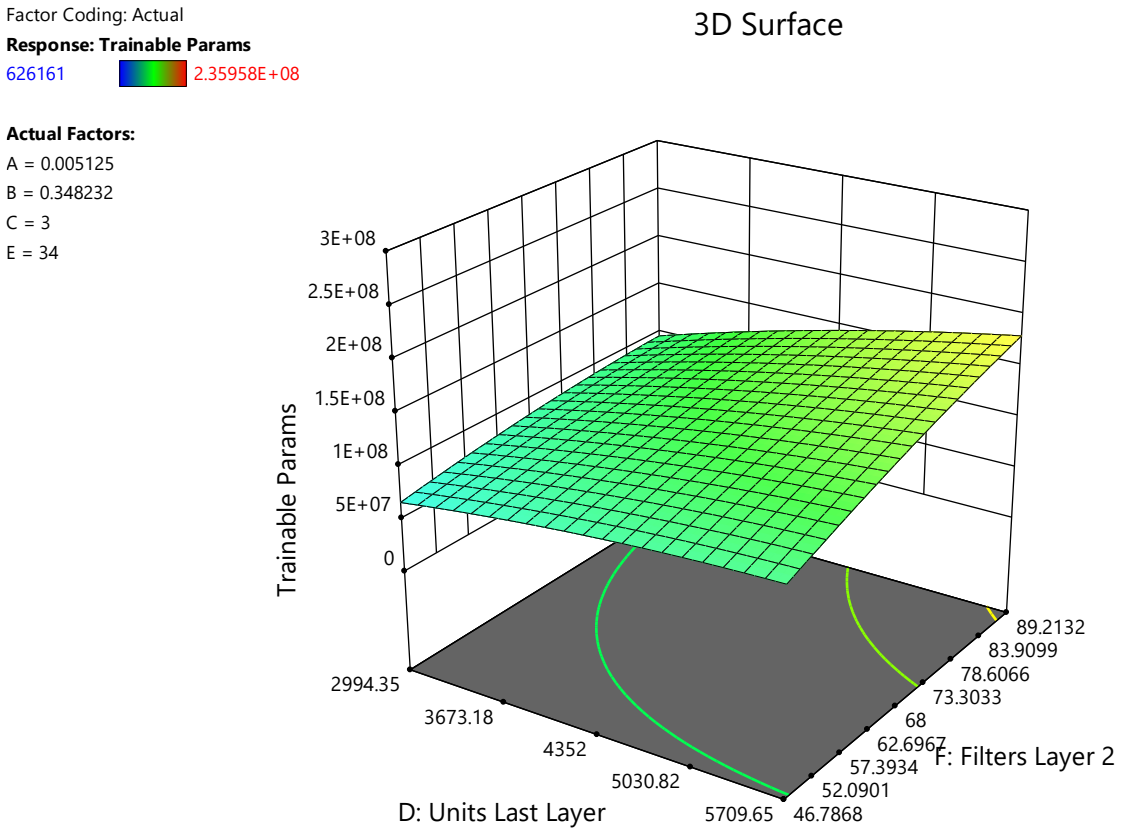


Figure 12. 3D Surface Plot of Trainable Parameters vs. Units Before Last Layer and Filters in Second Layer

Figure 11 and Figure 12 focus on the interaction between units before the last layer and filters in the second layer. Here too, an increasing trend is evident, although the surface is slightly less steep than the previous interactions, suggesting a somewhat weaker but still impactful DF interaction, as captured in the regression equation.

Together, these visualizations (Figures 7–12) validate the model's key findings:

- Model complexity is primarily driven by the architectural parameters (Conv2D layers, units, filters),
- Interaction effects significantly amplify complexity beyond simple additive behavior,
- Nonlinear (quadratic) terms are necessary for accurate prediction of trainable parameters.

1.3. DOE Analysis of Second Response: F1-Score Accuracy

For F1-score accuracy, the 2FI model again provides the best balance between fit and predictive ability, with an adjusted R^2 of 0.8199 and a predicted R^2 of 0.7158 (Table 3). ANOVA analysis, presented in Table 4, highlights learning rate (A), number of Conv2D layers (C), units before the last layer (D), and filters in the first layer (E) as significant factors. Interaction effects AC, AD, and AE were also significant, reinforcing that hyperparameter combinations, not just individual factors, govern model performance.

The regression equation for F1-score accuracy is:

F1-Score ACC = +0.4662 − 0.1348A − 0.0298C − 0.0530D − 0.1026E − 0.0121AC − 0.0147AD − 0.0326AE − 0.0116DE

Learning rate remains the most influential single factor, and the significance of interaction terms suggests that hyperparameter tuning must be multivariate rather than independent.

Table 3. Model Selection for Second Response

Source	**Sequential p-value**	**Lack of Fit p-value**	**Adjusted R^2**	**Predicted R^2**	
Linear	< 0.0001	0.0023	0.5966	0.5435	
2FI	**< 0.0001**	**0.0105**	**0.8199**	**0.7158**	**Suggested**
Quadratic	0.5775	0.0100	0.8161	0.6427	
Cubic	0.0617	0.0175	0.8713		Aliased

Table 4. ANOVA Summary for First Response

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	1.12	8	0.1395	37.09	< 0.0001	**significant**
A-Learning Rate	0.5874	1	0.5874	156.12	< 0.0001	
C-Conv Layers	0.0537	1	0.0537	14.28	0.0003	
D-Units Last Layer	0.0674	1	0.0674	17.91	< 0.0001	
E-Filters Layer 1	0.2522	1	0.2522	67.04	< 0.0001	
AC	0.0343	1	0.0343	9.10	0.0035	
AD	0.0343	1	0.0343	9.12	0.0035	
AE	0.1694	1	0.1694	45.03	< 0.0001	
DE	0.0146	1	0.0146	3.88	0.0528	
Residual	0.2709	72	0.0038			
Lack of Fit	0.2699	68	0.0040	15.88	0.0075	**insignificant**
Pure Error	0.0010	4	0.0003			
Cor Total	1.39	80				

Moving to the second response, F1-score accuracy, diagnostic plots were again assessed to verify model adequacy. Figure 13 presents the normal probability plot of residuals for F1-score accuracy. The residuals align closely with the reference line, confirming that the normality assumption is satisfied with this model as well. Figure 14 shows the residuals versus run order plot. Residuals are scattered randomly, indicating independence and absence of time-related systematic error, further supporting the model validity.

The effects of hyperparameters on F1-score accuracy were then visualized through a series of contour and 3D surface plots:

Figure 15 and Figure 16 explore the relationship between learning rate and Conv2D layers. Both plots reveal an optimal region of moderate learning rate (0.0005–0.005) and three Conv2D layers, where F1-score accuracy peaks. This matches the findings of the DOE scatterplot earlier (Figure 2) and statistically significant factors A (learning rate) and C (Conv2D layers) from the ANOVA results.

Figure 17 and Figure 18 display the interaction between learning rate and units before the last layer. The plots show that while increasing units can improve performance, it only does so if coupled with an optimal learning rate; otherwise, performance deteriorates. This validates the significant AD interaction observed in Table 4. Figure 19 and Figure 20 investigate the relationship between learning rate and filters in the first layer. Both plots highlight that the first convolutional layer's filter size needs to be balanced carefully with learning rate: too many filters or too aggressive learning rates degrade the F1-score.

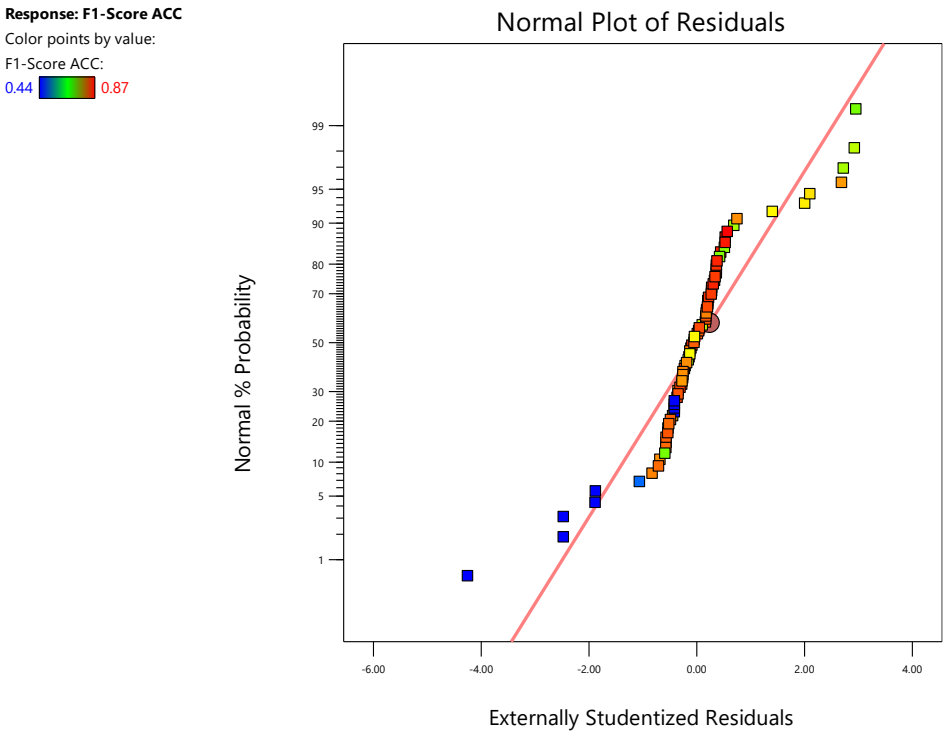


Figure 13. Normal Probability Plot of Residuals for F1-Score Accuracy

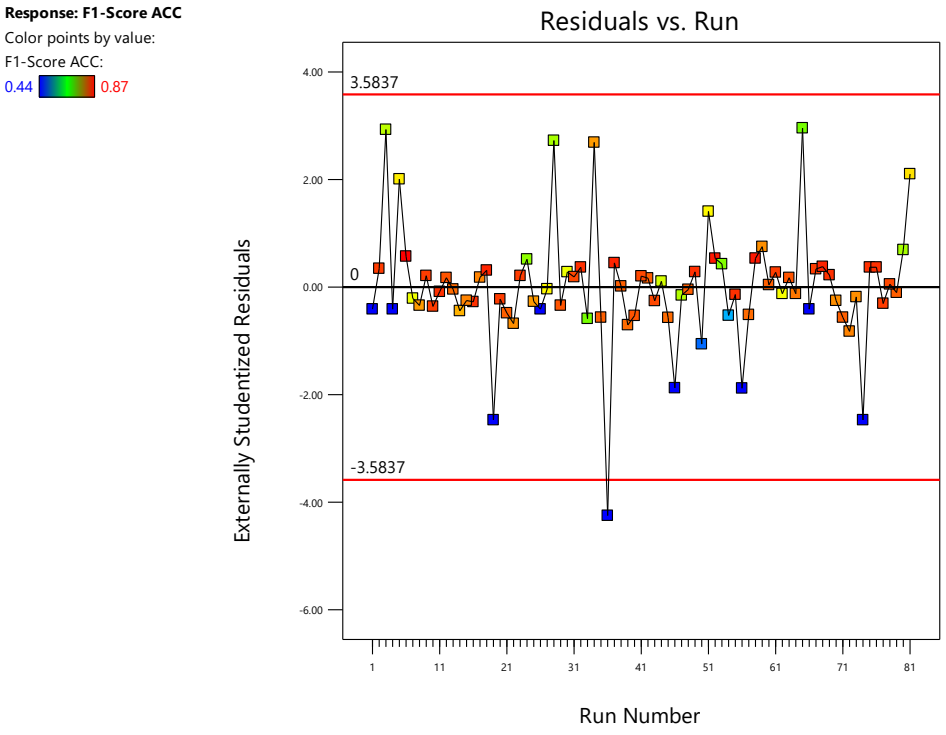


Figure 14. Residuals vs. Run Order for F1-Score Accuracy

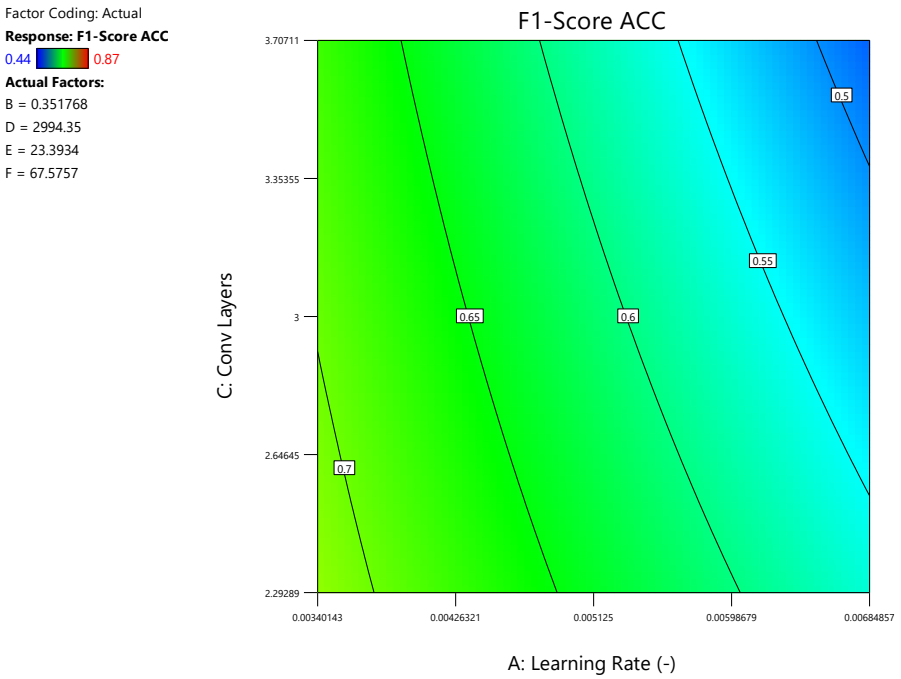


Figure 15. Contour Plot of F1-Score Accuracy vs. Learning Rate and Conv2D Layers

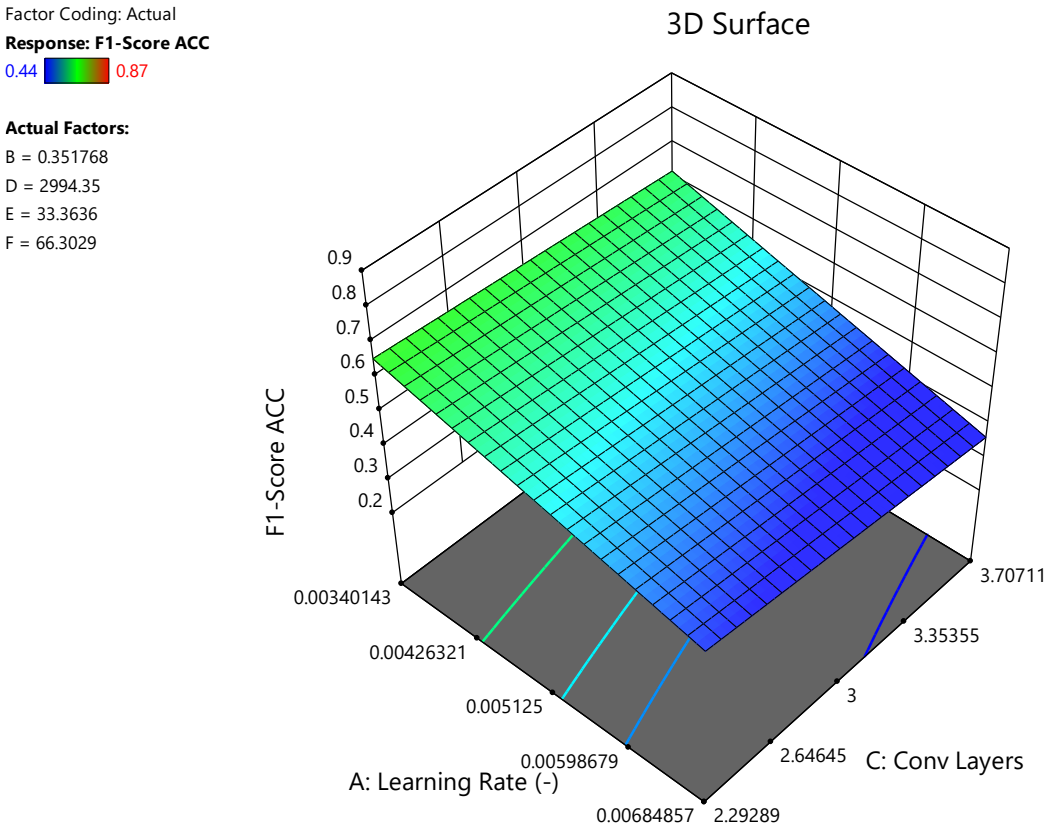


Figure 16. 3D Surface Plot of F1-Score Accuracy vs. Learning Rate and Conv2D Layers

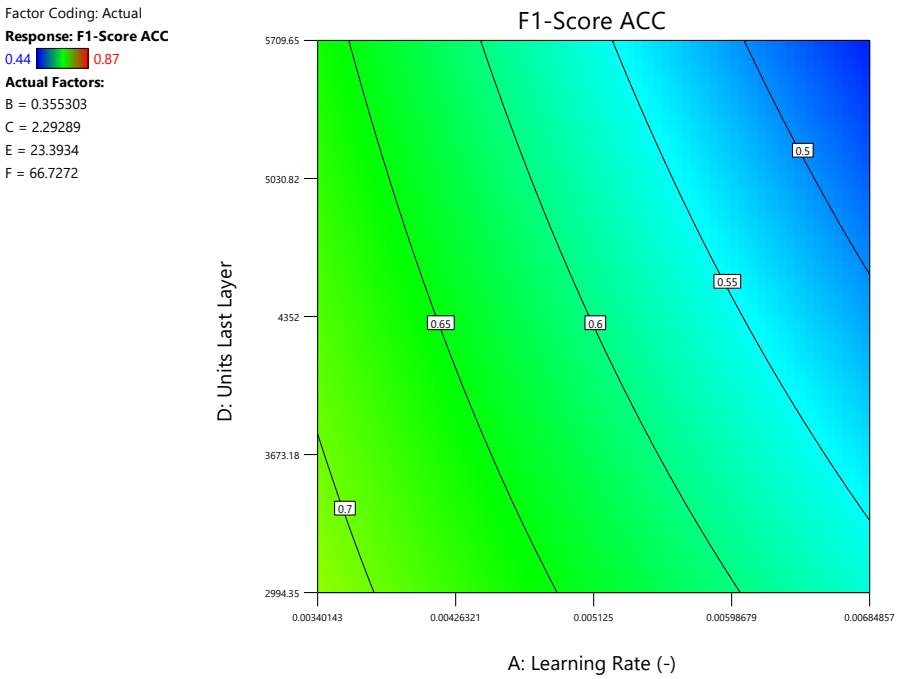


Figure 17. Contour Plot of F1-Score Accuracy vs. Learning Rate and Units Before Last Layer

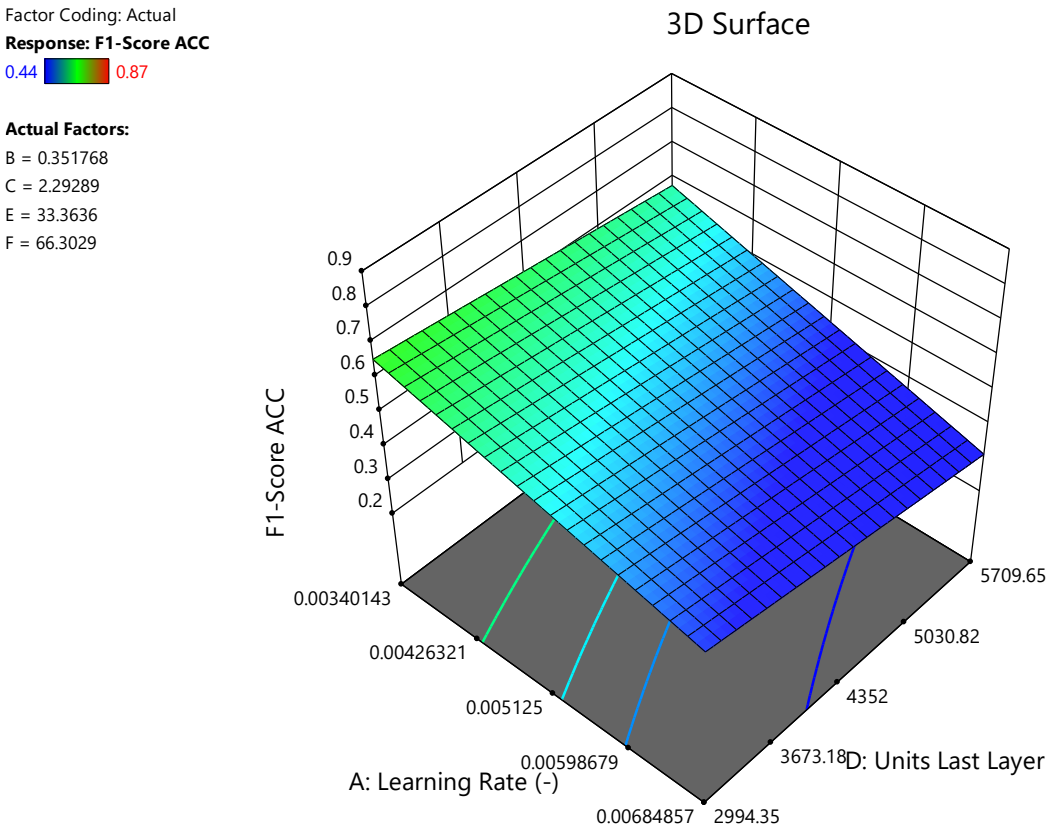


Figure 18. 3D Surface Plot of F1-Score Accuracy vs. Learning Rate and Units Before Last Layer

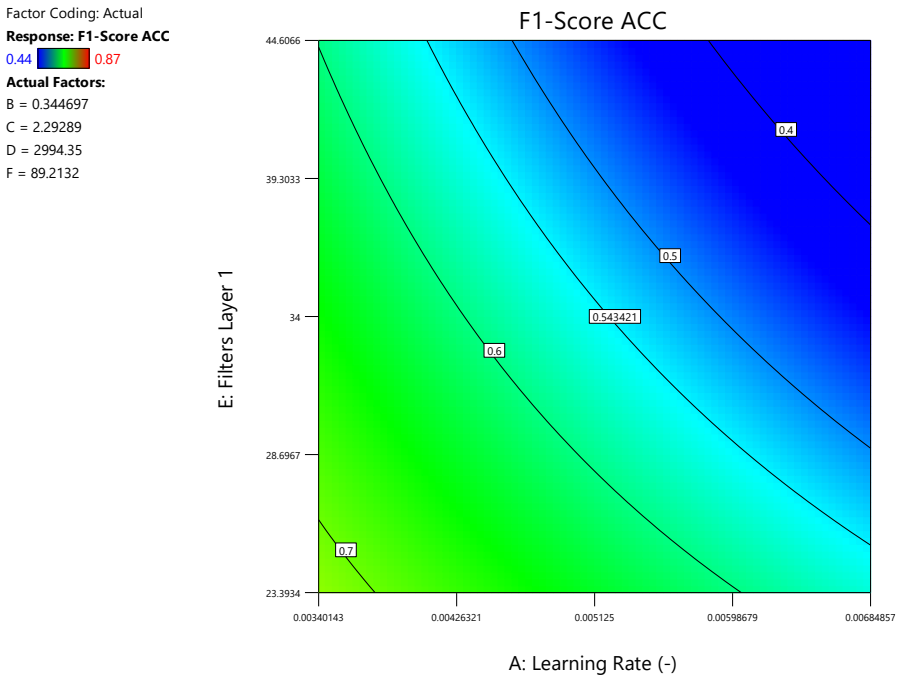


Figure 19. Contour Plot of F1-Score Accuracy vs. Learning Rate and Filters in First Layer

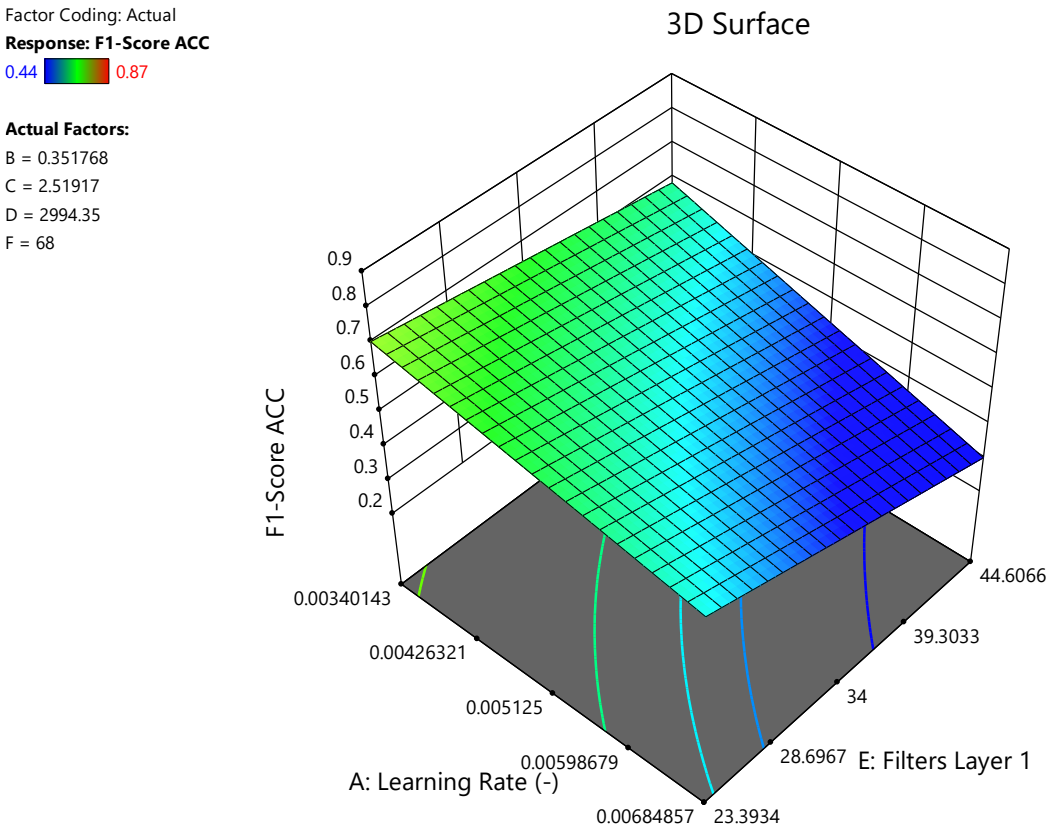


Figure 20. 3D Surface Plot of F1-Score Accuracy vs. Learning Rate and Filters in First Layer

These results correspond with the significant AE interaction term from the statistical model. Thus, Figures 15–20 consistently illustrate that:

- Learning rate is the most critical individual factor impacting F1-score,
- However, interactions between learning rate and architectural choices (Conv2D layers, units, filters) must be considered jointly,
- Optimal F1-score is achievable only when multiple hyperparameters are tuned together, not in isolation.

The post-DOE fine-tuning phase discussed in section 3.

2. CNN-Specific Insights

Model Complexity vs. Performance

- Increasing Conv2D layers and filter sizes does increase trainable parameters dramatically but does not always translate to better performance.
- Over-parameterization can lead to overfitting or unstable convergence. Numerical values are also reported in attached Excell sheet in the result section of the project repository in GitHub.

Learning Rate is Crucial

- A learning rate of 0.0005–0.005 performed best. Both underfitting (too low) and divergence (too high) occurred outside this range.
- Strong interactions between learning rate and units or filters suggest that adaptive tuning (e.g., LR-Scheduler) is essential. Such advanced control was also investigated in this comprehensive study, and the results have been presented in the result section of the GitHub repository.

Recommendations for CNN Tuning (Based on DOE)

- DOE models can be employed to predict optimal configurations: e.g., plug into regression equations for predictions.
- Avoid unnecessary architectural growth: deeper isn't always better.
- Always validate with multiple epochs: performance can evolve at different epochs. → In this respect, further optimization using advanced control systems have been developed analyzing the CNN results by finding the best epochs employing the validation accuracy early stopping function. Results of this study are also presented in the results section of the GitHub repository.

3. Post-DOE Fine Optimization of Selected Runs

Following the DOE, 22 runs exhibiting a favorable trade-off between lower trainable parameters and higher F1-score accuracy were selected for advanced fine-tuning (runs are presented in Figure 21). Optimization strategies included implementing learning rate decay schedules and early stopping based on validation accuracy to find the best epoch in each of the selected 22 runs. As the results show, this post-processing phase successfully improved the F1-score further, confirming the crucial role of dynamic learning rate control and careful early termination of training. According

to Figure 21, the Post-DOE with advanced control systems improved both average- as well as maximum F1-score accuracy values from 76.38 and 87.00% to 88.59 and 93.00%, which is a very clear improvement. According to the reported Excell sheet uploaded in the result section of the GitHub repository, in the post-DOE study, except run number 9 with an uncertain condition, overfitting was controlled successfully in all the other 21 runs, which is an indication for the success of the advanced controlling system via implementing learning rate decay schedules and early stopping based on validation accuracy to find the best epoch in terms of F1-score accuracy with no overfitting issue.

The results of this subsequent optimization process support the initial conclusions of the DOE. The learning rate was constantly present as the most important factor affecting performance, however, the architectural parameters need to be optimally chosen not to be too complex and not to lose in the accuracy. More details can be found in the Excel sheet "Results of Post-DOE" from the GitHub repository of this project, with highest F1-score accuracy obtained.

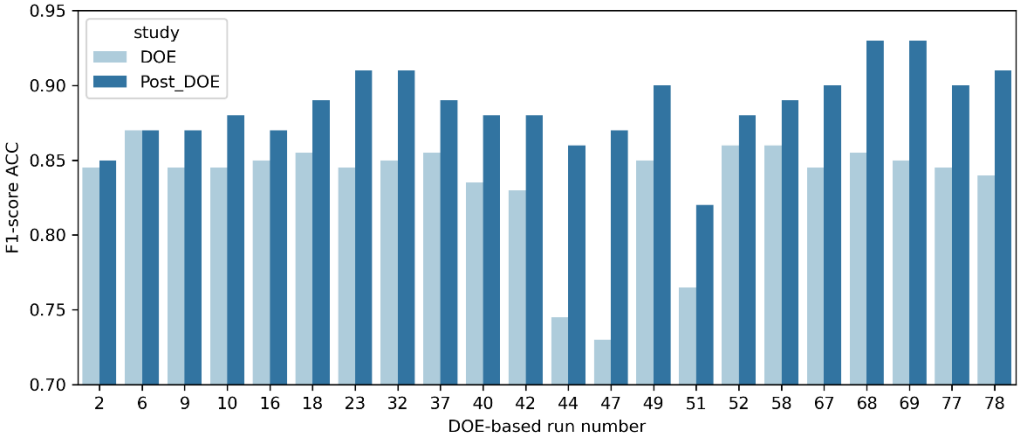


Figure 21. Results of Post-DOE Compared to DOE Studies in Selected DOE-Based Run Numbers

4. Summary & Conclusions

This work proves the effectiveness in the use of the structured DOE approach along with the optimization of the deep learning model. The CCD method served as a statistically rigorous model to comprehend the effects of hyperparameters on CNN's gender classification performance. Via sequential modeling and analysis, the learning rate was revealed to be the most important hyperparameter and the interaction effect between the learning rate and the architectural aspects were heavily confirmed.

Post-DOE optimization using callbacks such as learning rate scheduling and early stopping further enhanced the results, achieving high-performance models with manageable complexity and advanced controlled fitting condition.

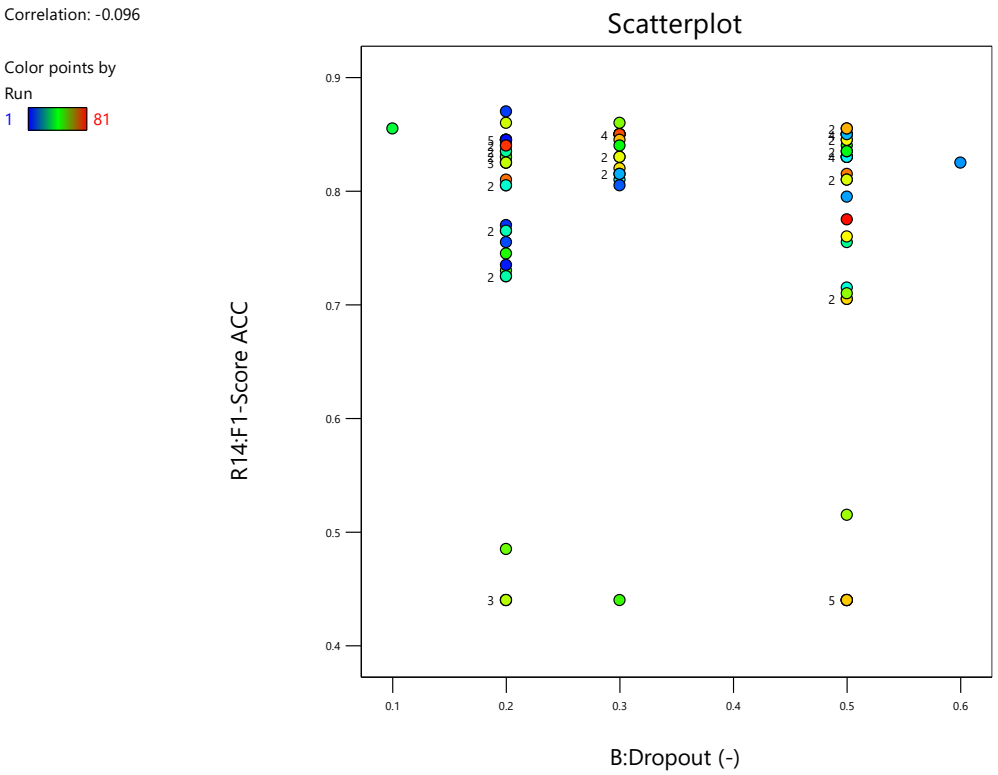
On the whole, the combination of DOE with deep learning is a promising method for systematic and efficient model tuning, making well-informed decisions beyond conventional manual or grid search techniques.

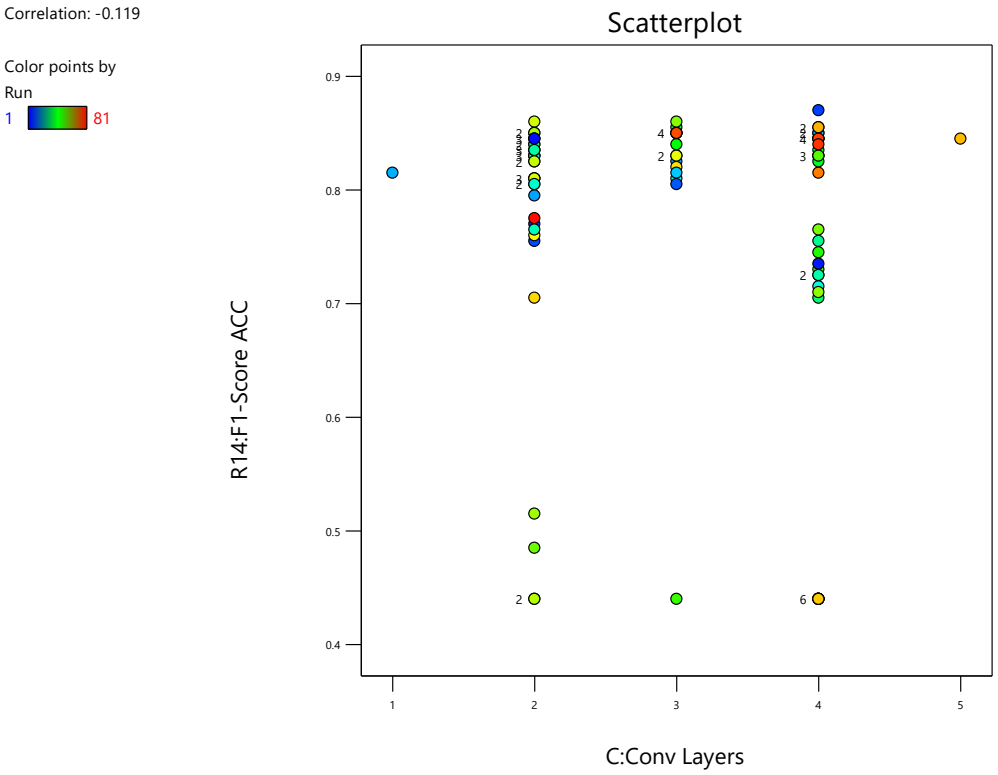
To further optimize and improve the performance based on F1-score accuracy of the CNN model applied, the best post-DOE model with the highest hyperparameters could be attached to each other as a single big CNN/ResNet50 image classification model.

Furthermore, as a future study, the selected model with optimized hyperparameter settings can be validated using a larger external dataset, since the dataset employed in this study was relatively small due to data availability constraints and limited local computational resources.

5. Appendix

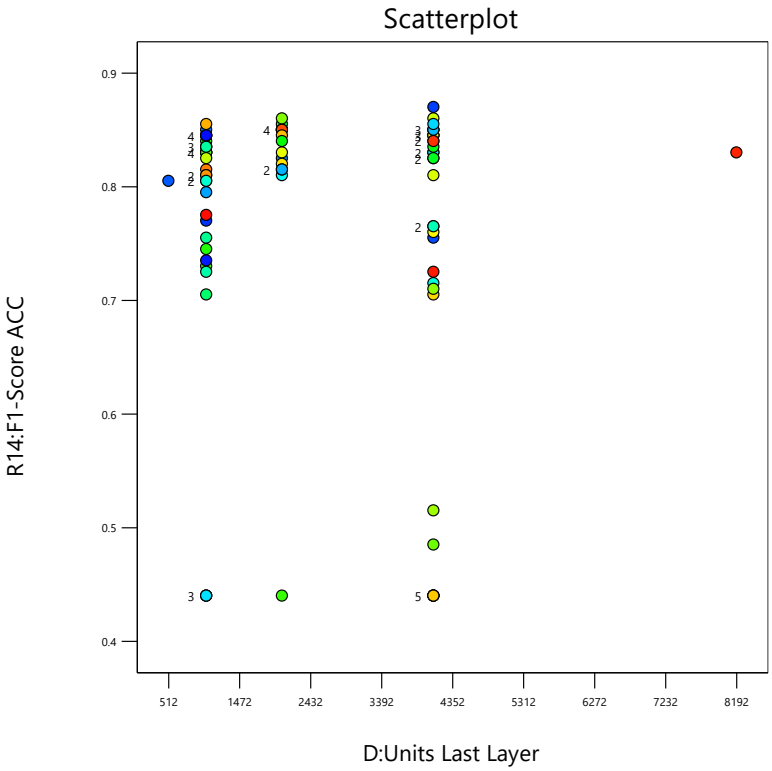
The appendix scatterplots, which are shown in Figure A1, further illustrate pairwise relationships between various hyperparameters and response metrics. These supplementary visuals confirm patterns observed in the main analysis, such as the importance of learning rate tuning and the exponential growth of trainable parameters with deeper or wider CNN architectures. They serve to reinforce the DOE conclusions without introducing new contradictory trends.





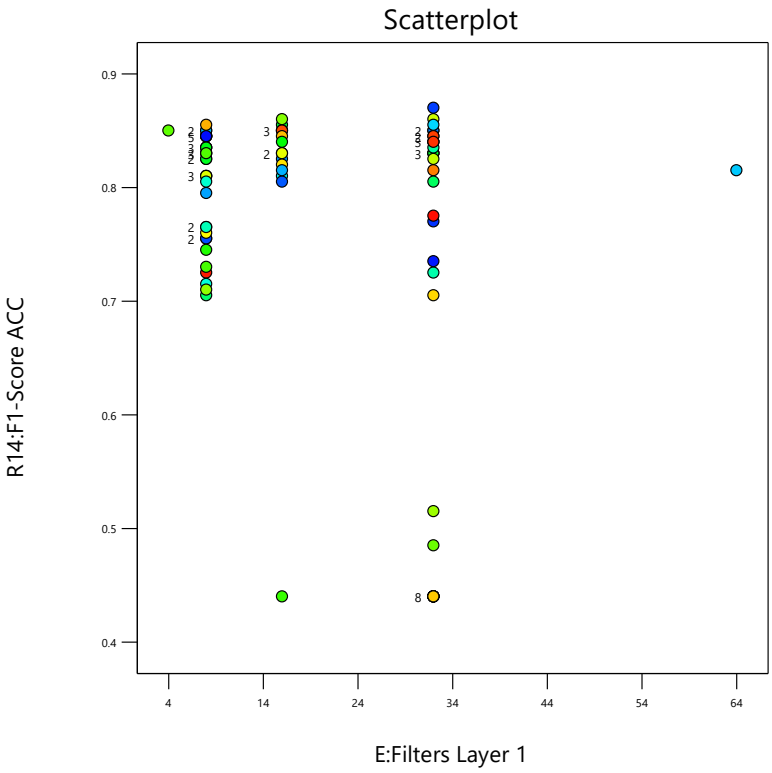
Correlation: -0.124

Color points by
Run
1 81



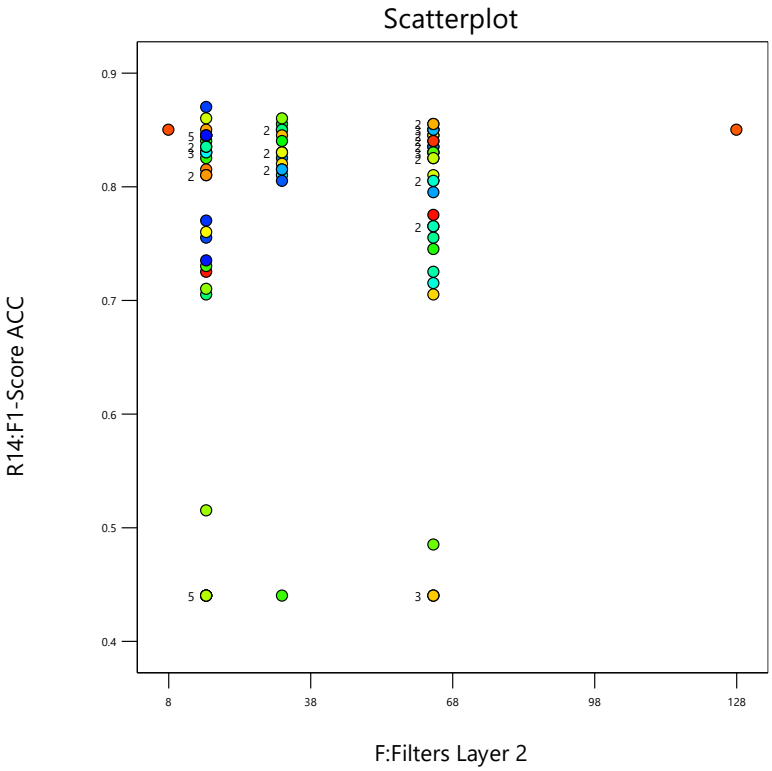
Correlation: -0.306

Color points by
Run
1 81



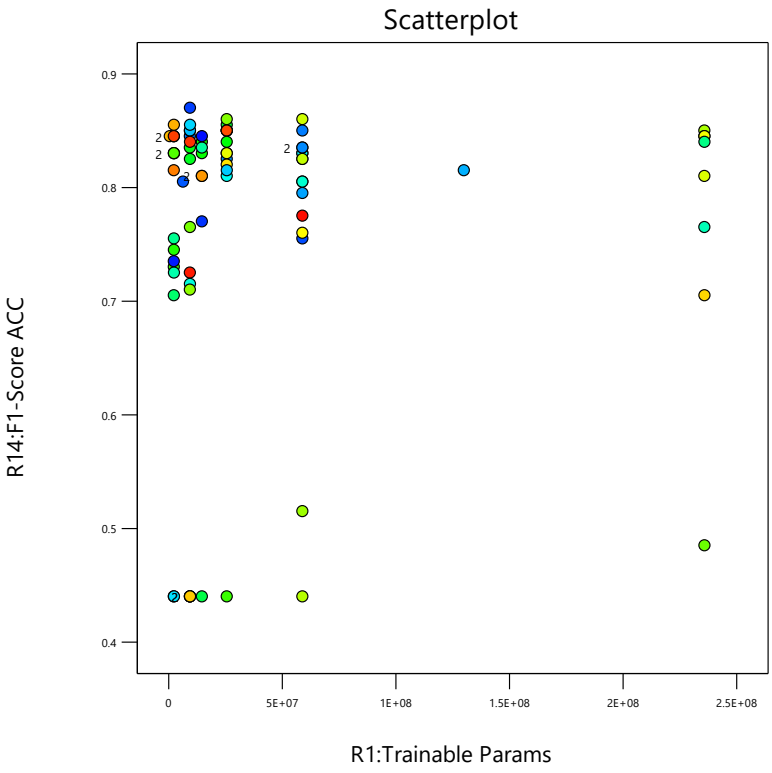
Correlation: 0.069

Color points by
Run
1 81



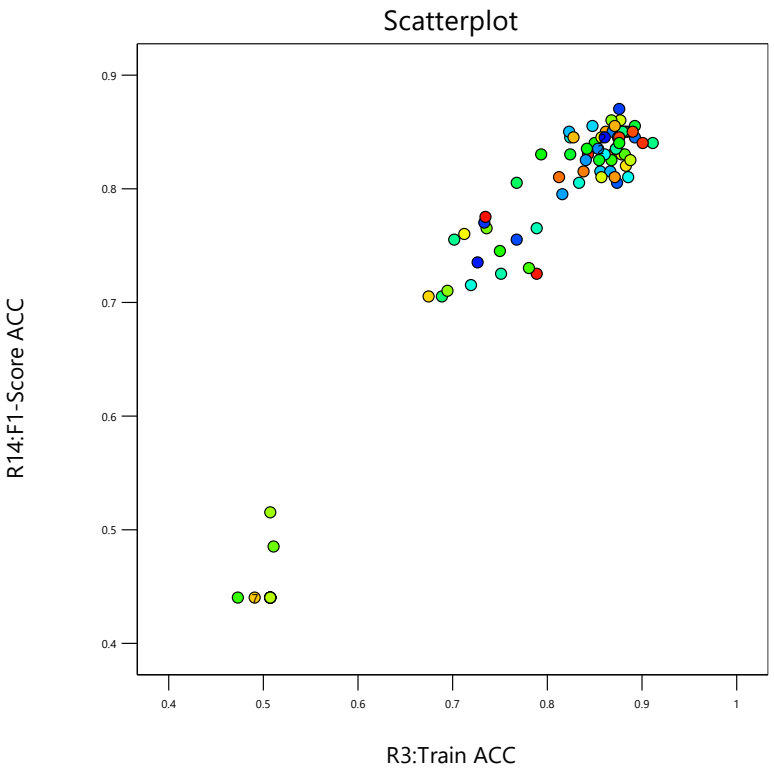
Correlation: 0.044

Color points by
Run
1 81



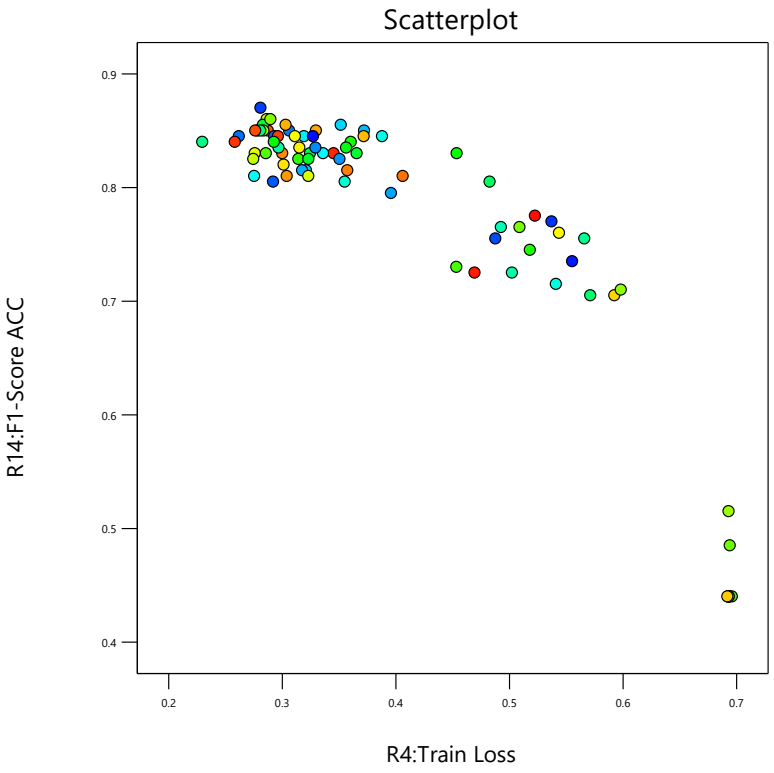
Correlation: 0.970

Color points by
Run
1 81



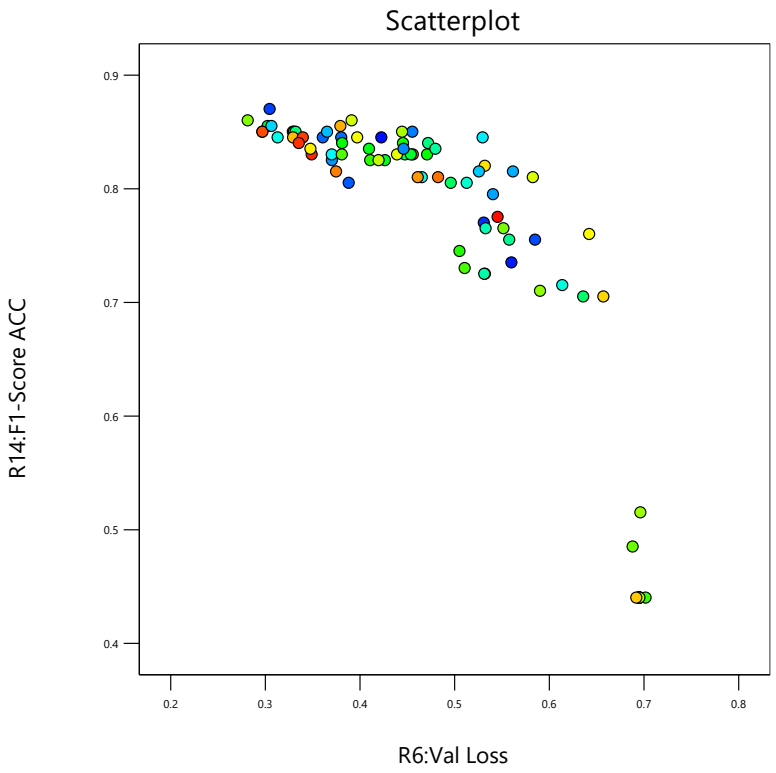
Correlation: -0.905

Color points by
Run
1 81



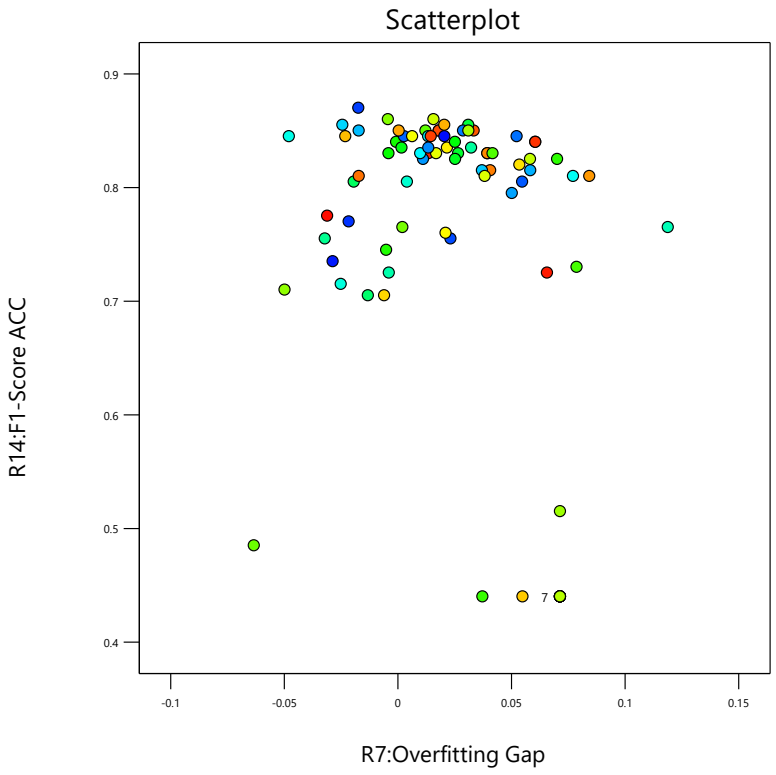
Correlation: -0.836

Color points by
Run
1 81



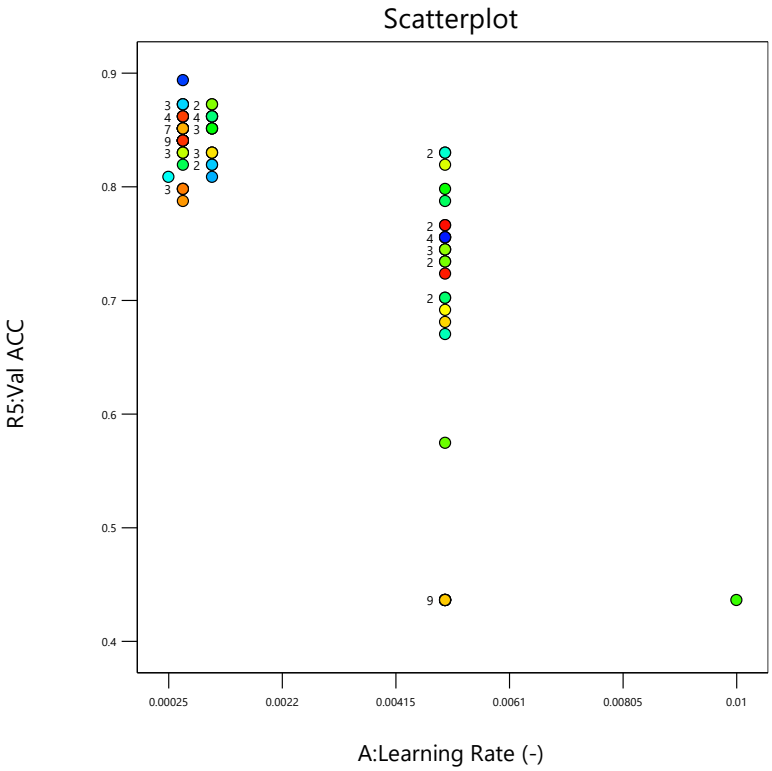
Correlation: -0.304

Color points by
Run
1 81



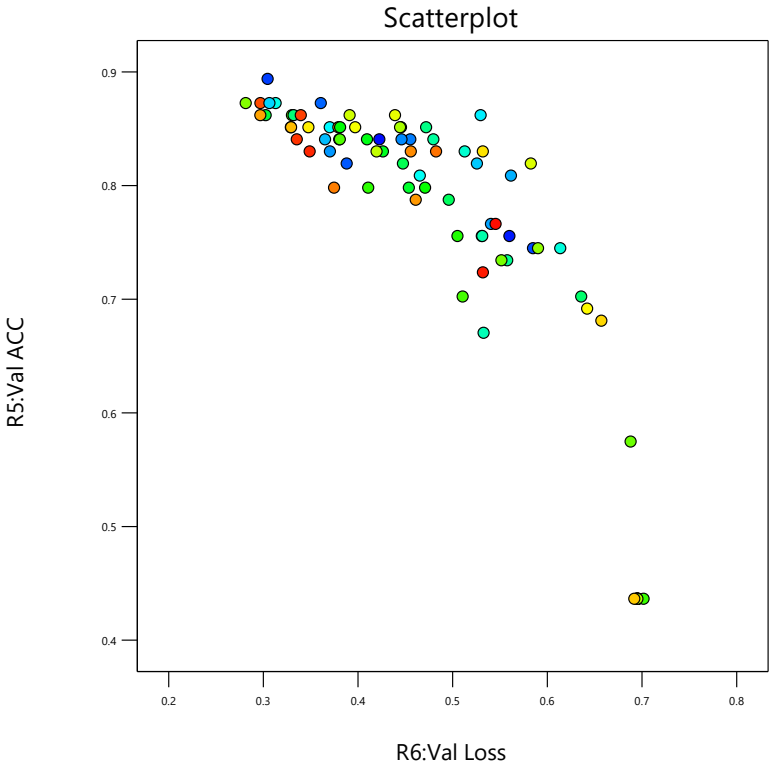
Correlation: -0.722

Color points by
Run
1 81



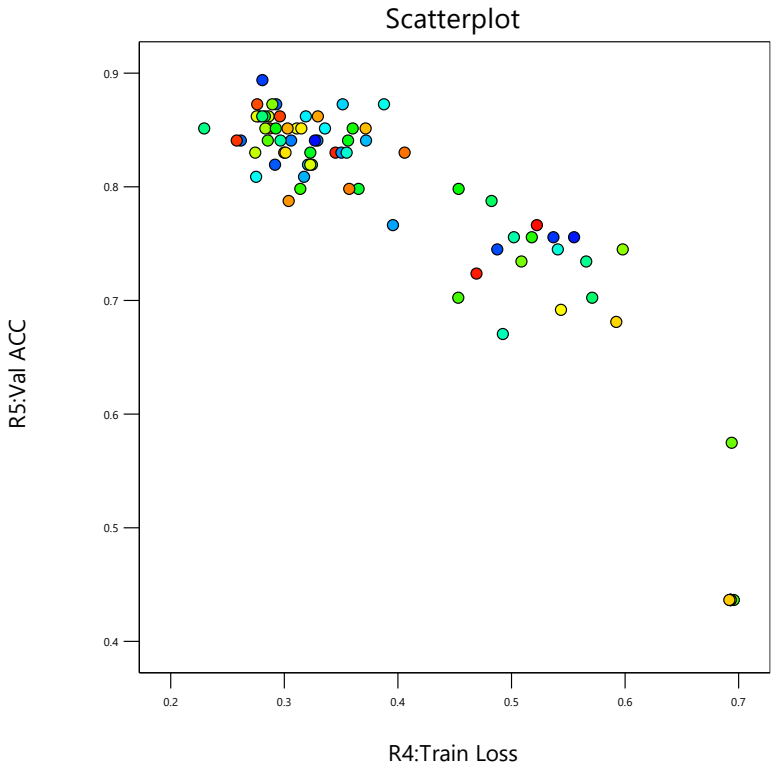
Correlation: -0.851

Color points by
Run
1 81



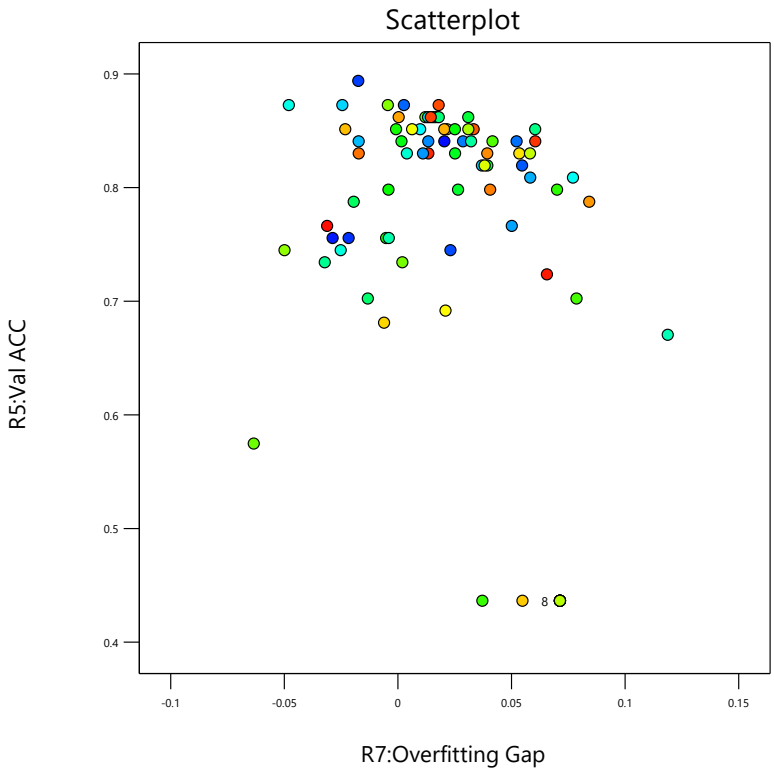
Correlation: -0.911

Color points by
Run
1 81



Correlation: -0.382

Color points by
Run
1 81



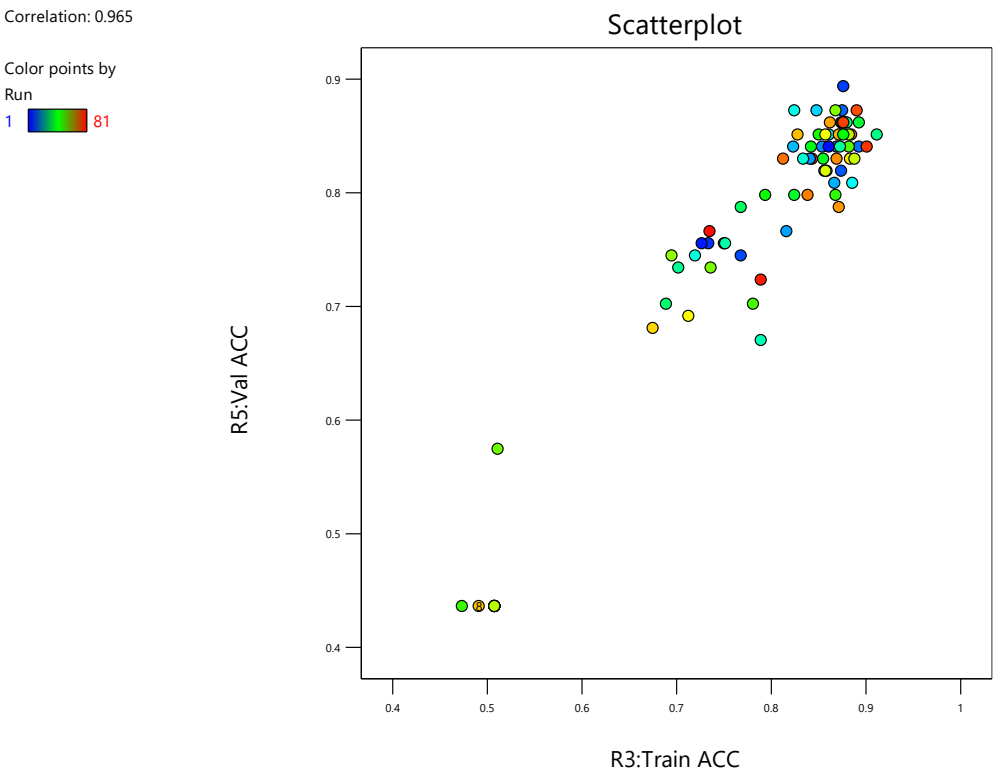


Figure A1. Supplementary Scatterplots of Hyperparameter & Responses Relationships