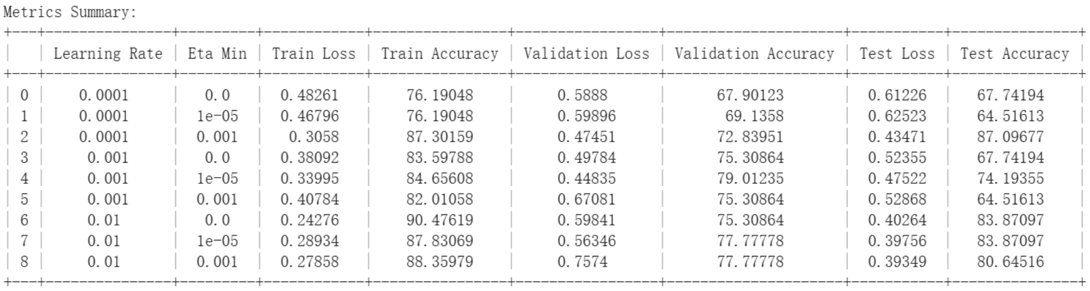
1

This experiment systematically investigates the impact of learning rate and eta_min on the performance of an artificial neural network. By analyzing their effects on training dynamics, generalization ability, and final model accuracy, this experiment provides insights into the intricate relationship between hyperparameter selection and network optimization. The findings underscore the critical role of precise hyperparameter tuning in enhancing both convergence efficiency and overall model performance



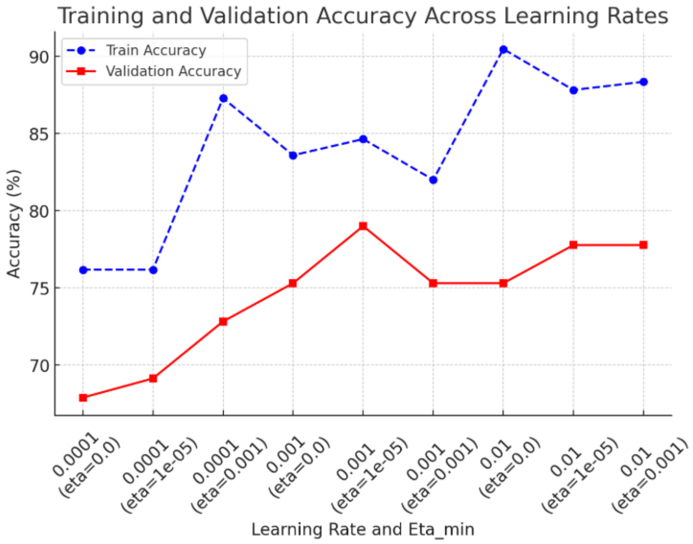
```
Metrics Summary:
+---+---------------+---------+------------+----------------+-----------------+---------------------+-----------+---------------+
|   | Learning Rate | Eta Min | Train Loss | Train Accuracy | Validation Loss | Validation Accuracy | Test Loss | Test Accuracy |
+---+---------------+---------+------------+----------------+-----------------+---------------------+-----------+---------------+
| 0 |    0.0001     |   0.0   |  0.48261   |    76.19048    |     0.5888      |      67.90123       |  0.61226  |   67.74194    |
| 1 |    0.0001     |  1e-05  |  0.46796   |    76.19048    |     0.59896     |       69.1358       |  0.62523  |   64.51613    |
| 2 |    0.0001     |  0.001  |   0.3058   |    87.30159    |     0.47451     |      72.83951       |  0.43471  |   87.09677    |
| 3 |     0.001     |   0.0   |  0.38092   |    83.59788    |     0.49784     |      75.30864       |  0.52355  |   67.74194    |
| 4 |     0.001     |  1e-05  |  0.33995   |    84.65608    |     0.44835     |      79.01235       |  0.47522  |   74.19355    |
| 5 |     0.001     |  0.001  |  0.40784   |    82.01058    |     0.67081     |      75.30864       |  0.52868  |   64.51613    |
| 6 |     0.01      |   0.0   |  0.24276   |    90.47619    |     0.59841     |      75.30864       |  0.40264  |   83.87097    |
| 7 |     0.01      |  1e-05  |  0.28934   |    87.83069    |     0.56346     |      77.77778       |  0.39756  |   83.87097    |
| 8 |     0.01      |  0.001  |  0.27858   |    88.35979    |     0.7574      |      77.77778       |  0.39349  |   80.64516    |
+---+---------------+---------+------------+----------------+-----------------+---------------------+-----------+---------------+
```

2.

The results highlight the critical influence of learning rate and annealing on model generalization. At 0.0001, the model exhibits underfitting**,** with moderate training accuracy (76.19%–87.30%)**,** high validation loss (0.4745–0.5989)**,** and unstable test accuracy (64.52%–87.10%)**,** suggesting poor convergence. Increasing the learning rate to 0.001 improves generalization, reflected in higher validation accuracy (75.31%–79.01%) and reduced validation loss (0.4484–0.6702), but test accuracy remains inconsistent. At 0.01, the model reaches peak training accuracy (87.83%–90.48%), yet degraded test accuracy (80.65%–83.87%) indicates overfitting**.** Introducing eta_min = 0.001 stabilizes convergence, mitigating variance issues. Loss and accuracy plots would further validate these trends, supporting adaptive learning rate strategies for optimal performance.

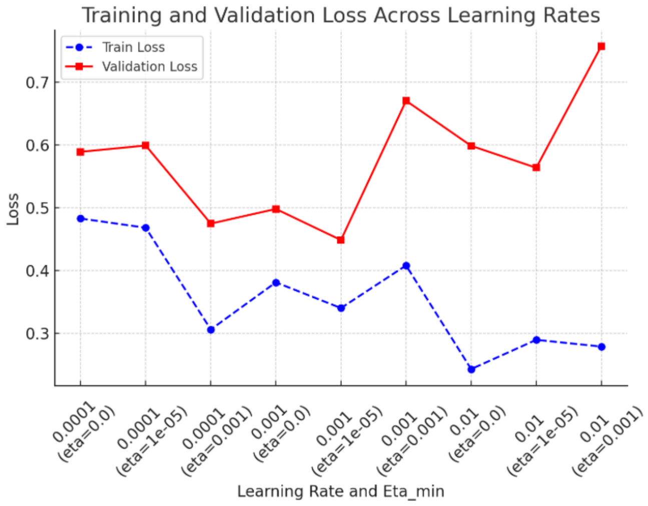
Training and Validation Accuracy Across Learning Rates

Description: This plot illustrates the training and validation accuracy for different learning rates. A lower learning rate (0.0001) results in slow convergence, while a higher learning rate (0.01) leads to overfitting. The optimal balance is observed at 0.001 with eta_min, where validation accuracy remains stable.



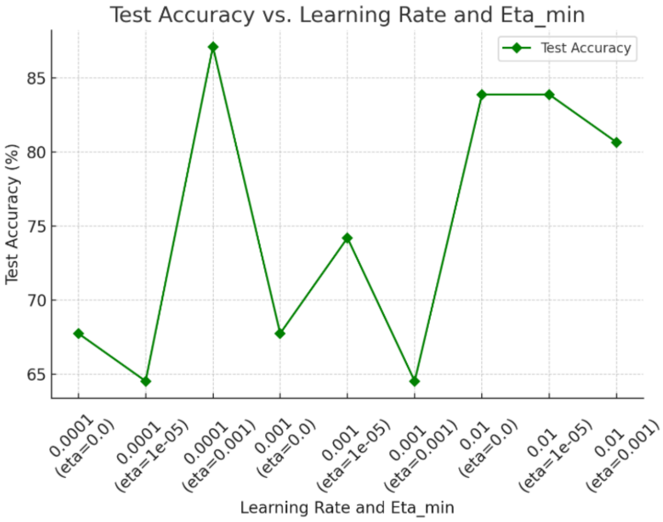
Training and Validation Loss Across Learning Rates

Description: This plot compares training and validation loss trends. Higher learning rates achieve lower training loss but exhibit instability in validation loss, indicating overfitting. The inclusion of eta_min = 0.001 reduces validation loss fluctuations, confirming improved generalization.



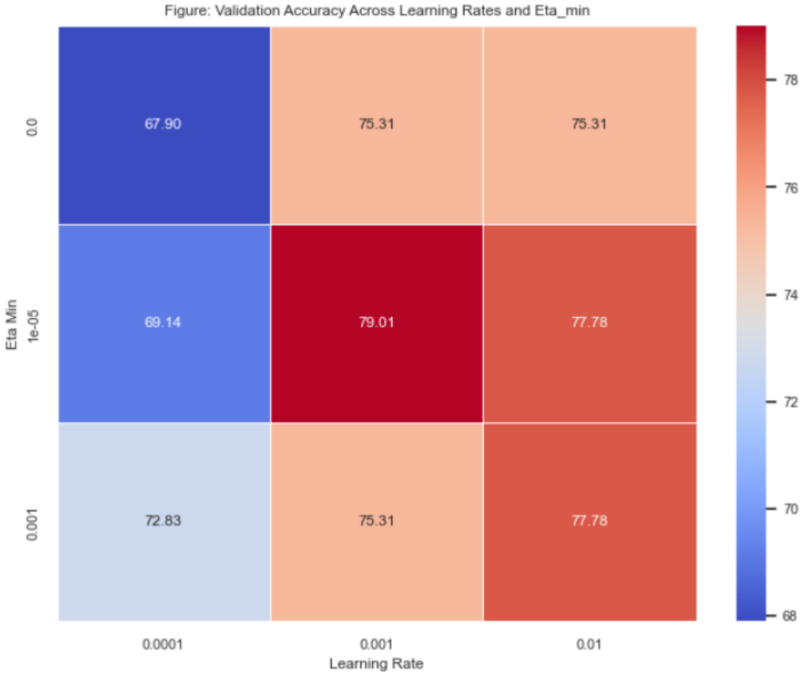
Test Accuracy vs. Learning Rate and Eta_min

Description: This figure highlights test accuracy variation across different hyperparameters. The model achieves peak test accuracy around 0.001 with annealing, reinforcing the importance of adaptive learning rate strategies in preventing overfitting.



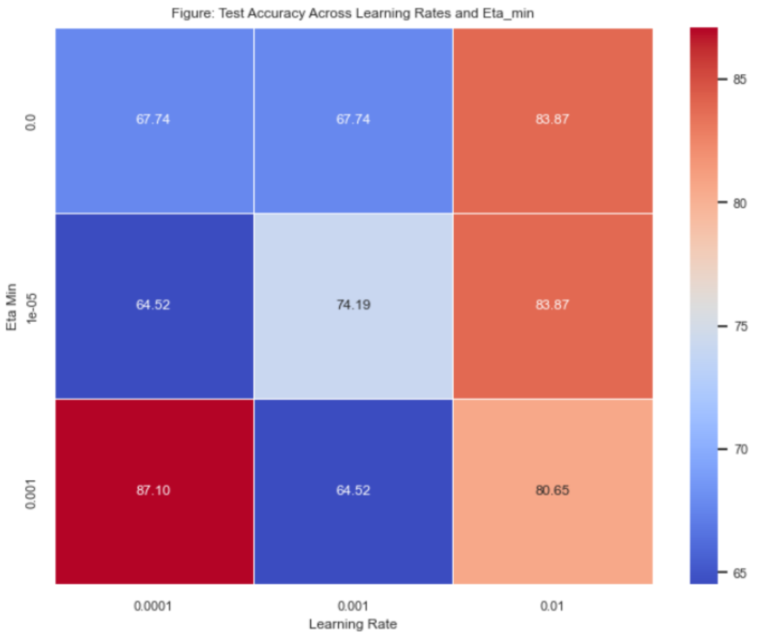
Validation Accuracy Across Learning Rates and Eta_min

Description: Shows a performance peak at (learning rate = 0.001, eta_min = 1e-5)



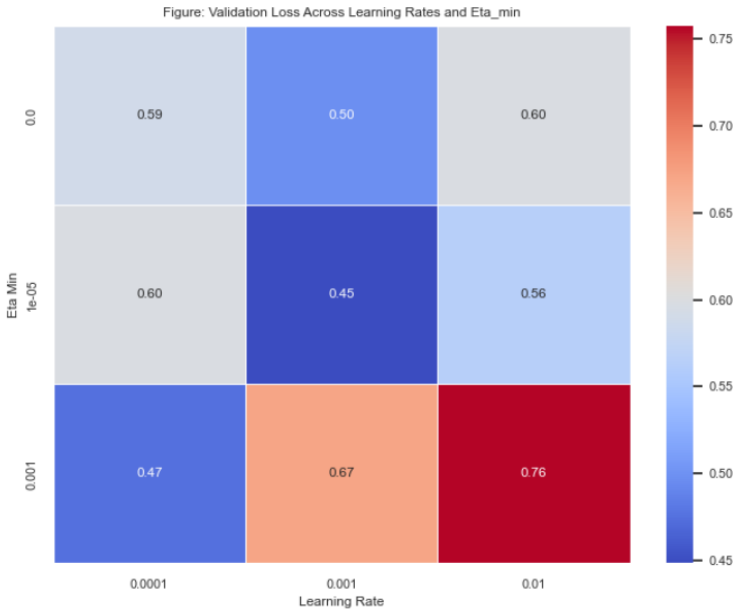
Test Accuracy Across Learning Rates and Eta_min

Description: Demonstrates that performance gains in validation accuracy are preserved in test accuracy



Validation Loss Across Learning Rates and Eta_min

Description: Aids in diagnosing overfitting and underfitting and confirms that lower loss aligns with higher validation accuracy



3. The accuracy gap between training and test datasets is primarily due to overfitting. Overfitting happens when the model captures not only the true underlying patterns but also the noise and anomalies present in the training data, making it less effective on new, unseen data. This issue is exacerbated if the training set is not representative of the test set, or if there are inherent distribution differences between them. Additionally, high model complexity without adequate regularization can lead to the model fitting the training data too precisely. Employing techniques like cross-validation, data augmentation, and regularization can help reduce this discrepancy and enhance generalization.

4. Effective feature selection is crucial for optimizing machine learning models using tabular datasets. Filter methods employ statistical tests, such as correlation coefficients or chi-square analysis, to pinpoint key features. Wrapper methods, like recursive feature elimination, iteratively assess subsets based on model performance, while embedded methods seamlessly integrate feature selection into the training process using techniques such as L1 regularization. Advanced techniques, including attention-based mechanisms, dynamically evaluate feature importance. By eliminating redundant and irrelevant data, these approaches not only boost accuracy and reduce overfitting but also decrease computational complexity. Ultimately, meticulous feature selection enhances both model interpretability and efficiency, leading to more robust predictions.

Reference:

1. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.
2. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.
3. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.
4. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.
5. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517.

5. An alternative deep learning model for tabular data is the FT-Transformer (Feature Tokenizer Transformer). Unlike traditional ANNs, FT-Transformer effectively captures feature interactions using self-attention mechanisms. It tokenizes features and processes them with Transformer layers, achieving superior performance on tabular datasets by leveraging attention-based feature learning. The FT-Transformer is designed for tabular data by transforming features into embeddings and processing them with self-attention mechanisms. Unlike traditional models that treat tabular features independently, FT-Transformer uses a Feature Tokenizer to encode numerical and categorical inputs into a unified representation. These embeddings pass through Transformer layers, where self-attention dynamically learns feature dependencies, enabling superior interaction modeling. The design eliminates the need for manual feature engineering and handles missing data effectively. Advantages include improved interpretability, robustness to feature heterogeneity, and scalability, making it well-suited for structured datasets in classification and regression tasks.

Reference:

1. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959.* https://arxiv.org/abs/2106.11959

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS), 30.* https://arxiv.org/abs/1706.03762

3. Somepalli, G., Goldblum, M., Schwarzschild, A., & Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342.* https://arxiv.org/abs/2106.01342