Predicting Gene Expression Response to Hyperoxia Using Machine Learning

Abstract: This project investigates the classification of gene regulation categories (Up-regulated, Down-regulated, No Change) using machine learning techniques. The dataset used consists of gene expression data under hyperoxia condition. As we implemented multiple classifiers, including Support Vector Machine (SVM), XG Boost, and K-Nearest Neighbors (KNN), and further improved performance through an ensemble model using Voting Classifier. The study highlights feature engineering techniques, model evaluation, and challenges encountered in achieving high predictive accuracy. The results indicate that while individual models perform adequately, the ensemble method provides a more balanced classification outcome.

Introduction: Gene regulation plays a crucial role in understanding biological responses to environmental conditions. In hyperoxia conditions, identifying genes that are up-regulated or down-regulated can provide insights into cellular adaptation mechanisms. However, due to the complexity of gene expression data, classification remains challenging due to the complexities in the data. This study employs machine learning techniques to classify gene regulation changes, emphasizing explainable modeling and feature selection to improve interpretability and performance. Prior studies in computational biology have leveraged machine learning to analyze gene expression data. Traditional statistical approaches, such as differential gene expression analysis, have limitations in capturing non-linear relationships in data. Machine learning models, particularly ensemble methods, provide an opportunity to improve predictive accuracy while retaining model explainability. This research builds upon previous methodologies by integrating feature engineering, hyperparameter optimization, and ensemble learning to address classification challenges in hyperoxia-induced gene expression changes.

Problem Statement: The classification of gene expression changes (Up-regulated, Down-regulated, No Change) under hyperoxia conditions is a challenging task due to the high dimensionality of the dataset and potential noise in the data. The primary goal of this project is to develop machine learning models that accurately classify gene regulation while ensuring interpretability and robustness. Hyperoxia is known to induce oxidative stress, leading to significant transcriptional changes in cells. Accurately classifying these changes is crucial for understanding cellular response mechanisms and identifying potential therapeutic targets. A robust classification model can facilitate biological discoveries and improve our understanding of gene-environment interactions.

Dataset and Analysis: The dataset used for this study is "CrePosHyperoxiavsCreNegHyperoxia_all_detectable_genes.csv," which contains gene expression values under hyperoxia conditions. The target variable (Regulation) is categorized based on fold change values:

- Up-regulated: Fold change > 1
- Down-regulated: Fold change < -1
- No Change: -1 ≤ Fold change ≤ 1

Features extracted include individual gene expression values, statistical metrics (mean, variance), and derived ratio-based features. Missing values were handled using mean imputation, and features were standardized using Standard Scaler.

Methodology:

1. **Feature Engineering:**
    - Calculation of CrePos Mean and Variance.
    - Creation of ratio-based features (Foldchange-to-Mean, p Value-to-Mean, FDR-to-Mean).
2. **Model Development:**

    Support Vector Machine (SVM): SVM is a powerful supervised machine learning algorithm for classification tasks. It works by finding the hyperplane that best separates the data into different classes. In your case, SVM is trained using a Radial Basis Function (RBF) kernel, which is effective for non-linear decision boundaries. The hyperparameters C (regularization parameter) and gamma (kernel coefficient) play crucial roles in determining the model's performance. Tuning these parameters through Grid SearchCV helps identify the optimal values by searching through a predefined range of values for each parameter. A higher C value puts more focus on reducing misclassification, while the gamma controls how far the influence of a single training example reaches. rained using an RBF kernel with hyperparameter tuning via Grid SearchCV to find optimal values of C and gamma.

    XG Boost: XG Boost is an optimized gradient boosting algorithm that has become popular due to its speed and accuracy, especially in binary classification tasks. In this case, XG Boost is fine-tuned for a binary classification problem (Up-regulated vs. Not Up-regulated) to classify gene expression changes. Hyperparameters like learning rate, max depth, and regularization factors are crucial to controlling the complexity and the speed of the model's learning process. The learning rate controls the contribution of each tree to the final prediction, max depth limits the depth of the trees (preventing overfitting), and regularization factors help balance model complexity and accuracy.

Hyperparameter tuning helps find the best combination of these factors for optimal performance. Optimized for binary classification (Up-regulated vs. Not Up-regulated) with hyperparameter tuning for parameters such as learning rate, max depth, and regularization factors.

K-Nearest Neighbors (KNN): KNN is a simple and intuitive algorithm that classifies data points based on the majority class of their nearest neighbors. For your task, KNN is used for multi-class classification, meaning it can classify gene expression changes into multiple categories: Up-regulated, Down-regulated, or No Change. The algorithm's sensitivity is influenced by the number of neighbors (k) used and the distance metric chosen (e.g., Euclidean or Manhattan distance). A small value of k might make the model sensitive to noise, while a larger value of k could smooth out the decision boundaries. By experimenting with different values of k and distance metrics, you can identify the optimal configuration for better accuracy. Implemented multi-class classification with varying numbers of neighbors and distance metrics to assess model sensitivity.

Ensemble Model (Voting Classifier): Ensemble learning involves combining multiple models to improve prediction accuracy by leveraging the strengths of each individual model. In this case, a Voting Classifier is used to combine SVM, XG Boost, and KNN. With soft voting, the predicted class is based on the weighted average of the probabilities predicted by each model. This allows the ensemble to make a final decision by considering the collective knowledge of all three models. The advantage of using an ensemble model is that it reduces the likelihood of overfitting, improves generalization, and enhances the robustness of the classifier by combining the different strengths of each model, leading to more reliable predictions. Combined SVM, XG Boost, and KNN using Voting Classifier (soft voting) to leverage strengths from multiple models for improved classification performance.

3. **Evaluation Metrics:**
    - F1-score (weighted)
    - Precision
    - Recall
    - Accuracy

Results: The models were evaluated on a separate test set. The following summarizes the best performance obtained:

- **SVM:** Best weighted F1-score after hyperparameter tuning.
- **XG Boost:** Achieved high precision in distinguishing Up-regulated genes with some overfitting which was resolved
- **KNN:** Provided a balanced multi-class classification performance.
- **Ensemble Model:** Outperformed individual models with an improved weighted F1-score and no overfitting.

Model	Best Parameters	F1-score (Weighted)
SVM	{'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}	0.94
XGBoost	{'n_estimators': 5, 'max_depth': 2, 'learning_rate': 0.001}	0.99
KNN	{'n_neighbors': 40, 'weights': 'distance'}	0.93
Ensemble	Combination of all three models	0.96

Conclusion: This study demonstrates that machine learning can be effectively used for gene regulation classification under hyperoxia conditions. While individual models achieved reasonable performance, the ensemble approach provided a more robust solution with no overfitting. Future work could explore deep learning methods, additional biological features, and more extensive datasets to further improve classification accuracy.

References:

1. https://figshare.com/articles/dataset/Gene_expression_csv_files/21861975
2. https://www.youtube.com/watch?v=oC7d1FQajjI
3. https://pmc.ncbi.nlm.nih.gov/articles/PMC8208445/