B. Tech. Project Report *Phase* I

" M17 media optimization using Random Forest Algorithm for efficient heparosan production by lactococcus lactis"

Submitted in partial fulfillment of requirements
for the award of the degree of Bachelor of Technology from IIT Guwahati

Under the supervision of
Prof. Senthilkumar Sivaprakasam

Submitted by
Rohit Raj Kohli
210106055

November, 2024
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati 781039, Assam, INDIA

Certificate

This is to certify that the work presented in the report entitled " **M17 media optimisation using Random Forest Algorithm for efficient heparosan production by lactococcus lactis**" by **Rohit Raj Kohli (210106055)**, represents an original work under the guidance of **Prof. Senthilkumar Sivaprakasam** at the Department of Biosciences and Bioengineering, Indian Institute of Technology, Guwahati. This study has not been submitted elsewhere for a degree.

Signature of student:

Rohit Kohli

Date: 11/11/2024
Place: IIT Guwahati

Rohit Raj Kohli
(210106055)

Signature of supervisor

15/11/2024

Date: 15/11/2024
Place: IIT Guwahati

Prof. Senthilkumar Sivaprakasam
Professor, Department of Biosciences and
Bioengineering, Indian Institute of
Technology, Guwahati

Signature of HOD

Date:
Place: IIT Guwahati

Head of
Department, Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati, India

ACKNOWLEDGEMENT

I would like to express my gratitude to the Indian Institute of Technology, Guwahati, for granting me the opportunity to pursue a Bachelor of Technology in Biosciences and Bioengineering. I want to specifically thank my project supervisor, Dr. Senthilkumar Sivaprakasam, a Professor at IIT Guwahati, for allowing me to undertake my bachelor's thesis project in the BioPAT laboratory. Throughout the project, his unwavering support, guidance, and supervision were invaluable.

I am also thankful to the Department of Biosciences and Bioengineering at IIT Guwahati for providing the necessary facilities for my research. Special thanks go to Mr. Tilak Raj S, a Ph.D. scholar at BioPAT, for his mentoring. Additionally, I want to express my sincere appreciation to my lab-mates, whose motivation and valuable feedback played a crucial role in the successful completion of my work. I extend my thanks and regards to my family for their constant support, well-wishes, and encouragement, which I consider priceless.

In conclusion, my heartfelt thanks go out to everyone who has been a part of my academic and research endeavors at IIT Guwahati. I am confident that the skills and knowledge gained during my undergraduate studies will serve as a solid foundation for my future endeavors in the field of biosciences and bioengineering.

Date: 11/11/2024

Rohit Raj Kohli
210106055

TABLE OF CONTENTS

ABSTRACT

This project investigates the application of machine learning techniques to optimize the medium composition for the growth of *Lactococcus lactis*, focusing on maximizing both final cell density and heparosan production. Through a comprehensive exploration of machine learning models, including linear models, ensemble methods, and deep neural networks, this research evaluates the suitability of each model for predicting complex, multi-dimensional relationships in bioprocess optimization. By rigorously comparing the performance of models such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), this study identifies the ANN as the most robust model for accurately capturing the non-linear dependencies within the dataset. The ANN model demonstrated superior predictive power, positioning it as the optimal tool for future applications in industrial biotechnology where growth media components play a pivotal role.

1. INTRODUCTION

1.1 Background

The optimization of growth media is essential in microbial biotechnology to enhance productivity, reduce costs, and maximize yield. Traditional methods for medium optimization are often time-consuming and resource-intensive, involving extensive trial-and-error experimentation. In the context of *Lactococcus lactis*, a bacterium widely used in food fermentation and industrial biotechnology, optimizing the medium composition directly impacts cell yield and product synthesis. In particular, the production of heparosan, a valuable precursor to heparin used in medical applications, is influenced by specific nutrients and environmental factors within the growth medium.

What is Heparosan?

- Heparosan is a polysaccharide produced by bacteria, including *Lactococcus lactis*, as a precursor to heparin. It is composed of repeating glucuronic acid and N-acetylglucosamine units and can be chemically modified to produce heparin.

Applications of Heparosan

1. **Heparin Production**:

    - Heparosan is converted to heparin, a widely used anticoagulant for preventing blood clots in surgeries and cardiovascular treatments.
2. **Drug Delivery and Biomaterials**:
    - Heparosan is used to create hydrogels and nanoparticles for controlled drug release, especially in targeted therapies.
3. **Tissue Engineering and Regenerative Medicine**:
    - It serves as a scaffold material in tissue engineering, aiding cell adhesion and supporting tissue growth, especially in wound healing.
4. **Antiviral and Antimicrobial Coatings**:
    - Heparosan has shown potential antiviral properties and is being explored as a coating on medical devices to prevent infections.

1.2 Role of Machine Learning in Bioprocess Optimization

Machine learning (ML) offers a data-driven solution for bioprocess optimization by identifying complex patterns within datasets, enabling researchers to predict optimal conditions without exhaustive experimentation. In this study, we aim to apply ML to identify the optimal composition of M17 media for *L. lactis* growth and heparosan production. ML models, ranging from simpler regression models to complex neural networks, are evaluated for their capacity to capture non-linear interactions among multiple features.

1.3 Problem Definition

The main objectives of this study are:

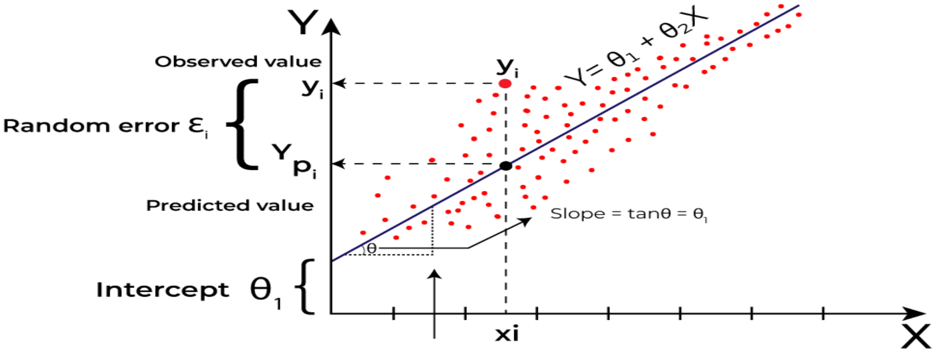
- To evaluate a range of machine learning models for their effectiveness in modelling the non-linear, high-dimensional relationships within growth media datasets.
- To select the best-performing model based on predictive accuracy and model robustness.
- To develop a machine learning pipeline that can predict optimal medium composition for maximizing both cell density and heparosan output in *L. lactis* cultures.

2. LITERATURE REVIEW

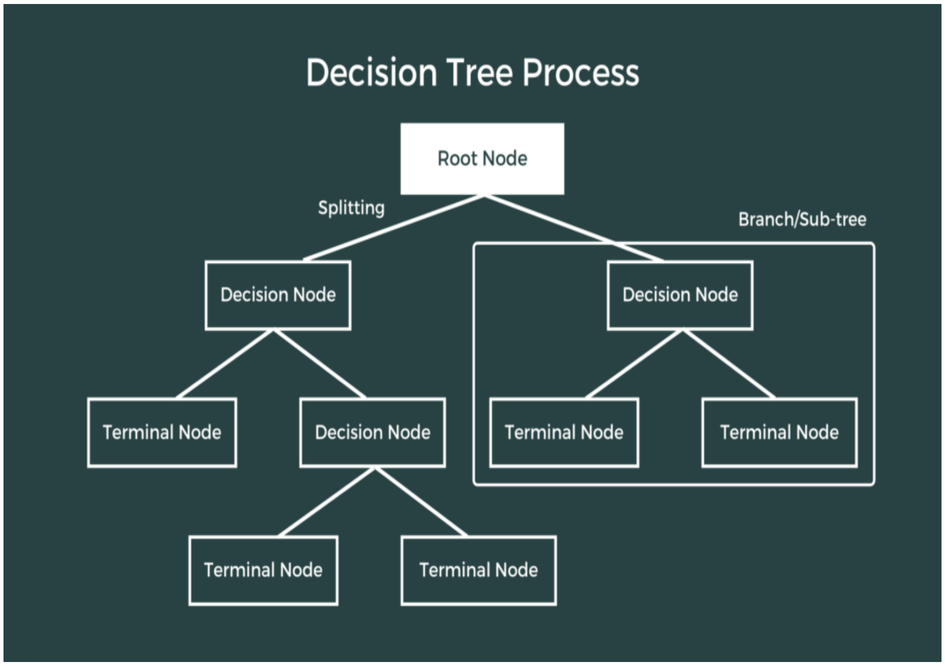
2.1 Overview of Machine Learning Models in Bioprocess Engineering

Machine learning models have been widely used in bioprocess engineering, especially in tasks requiring optimization of complex, multi-variable systems. Below is an in-depth exploration of various machine learning techniques and their applications:

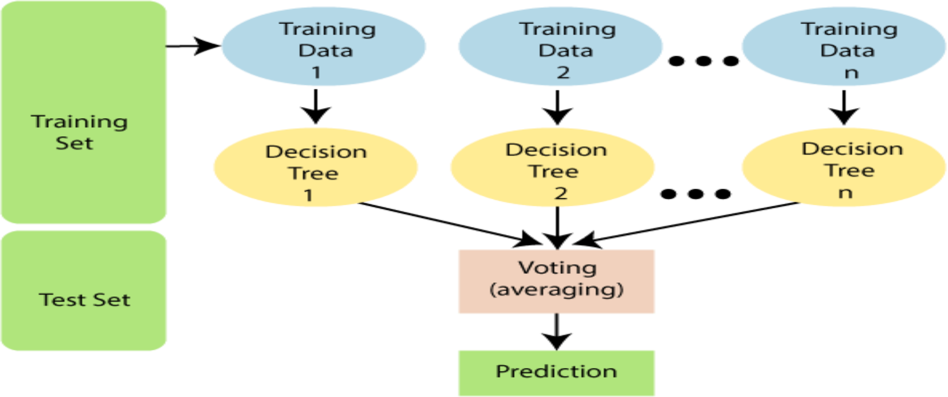
1. **Linear Regression and Polynomial Regression**:
    - *Application*: Linear regression models provide a foundational approach to understanding the relationships between independent and dependent variables. Polynomial regression extends this by fitting polynomial equations, capturing basic non-linearity.
    - *Limitations*: While useful as baseline models, linear regression and its polynomial counterparts struggle with high-dimensional data and complex biological interactions, often leading to oversimplified models that fail to generalize well in multi-dimensional tasks.


Y
Observed value
y_i
y_i
Random error ε_i
Y_{p_i}
Predicted value
$Y = \theta_1 + \theta_2 X$
Slope = $\tan\theta = \theta_1$
θ
Intercept θ_1
xi
X

2. **Decision Trees**:
    - *Application*: Decision Trees are tree-based models that segment data based on feature splits, offering interpretability and flexibility. They are commonly used in biological optimization tasks for their ease of use and interpretability.
    - *Advantages*: They are effective in non-linear relationships, can handle categorical data, and are less affected by feature scaling.
    - *Limitations*: Decision Trees are prone to overfitting, especially with small or unbalanced datasets, which limits their generalizability in complex bioprocess datasets.

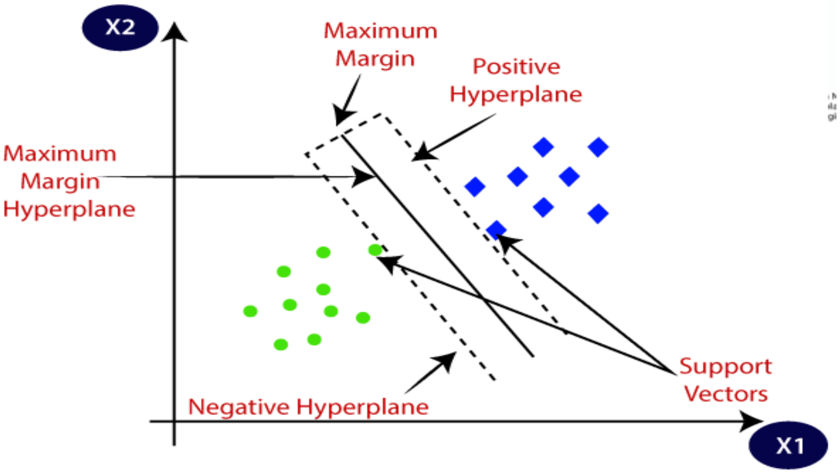


3. **Random Forests**:
   - *Application*: Random Forests, an ensemble of Decision Trees, address the overfitting problem by combining predictions from multiple trees. They are highly effective in handling noisy datasets typical of biological systems.
   - *Advantages*: High accuracy, reduced overfitting, and good feature importance insights.
   - *Limitations*: Random Forests can be computationally intensive with large datasets and may fail to capture complex interactions without tuning.

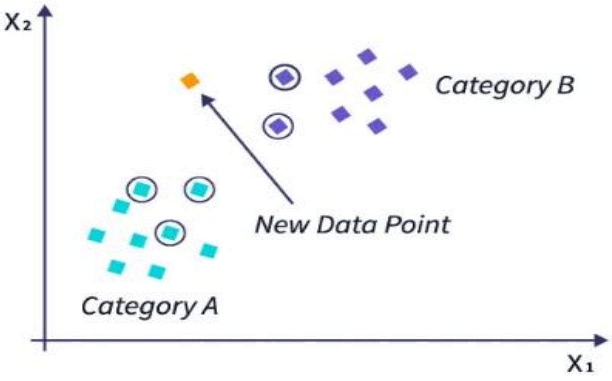


4. **Support Vector Machines (SVM)**:

- *Application*: SVM models are powerful for classification and regression tasks, particularly when the data is linearly separable in a high-dimensional feature space.
- *Advantages*: Effective in high-dimensional spaces and robust to overfitting when regularized.
- *Limitations*: SVMs struggle with large datasets and are sensitive to hyperparameter settings, making them less ideal for highly complex biological datasets.

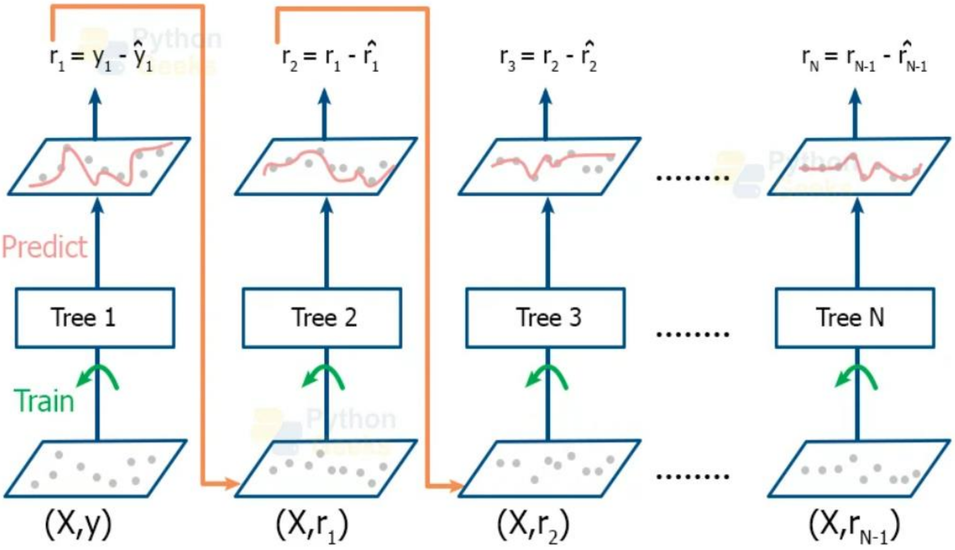


5. **K-Nearest Neighbours (KNN)**:
   - *Application*: KNN is a non-parametric model that predicts outcomes based on the nearest data points in feature space. It is suitable for low-dimensional tasks but has limited application in high-dimensional, complex data like bioprocess optimization.
   - *Limitations*: KNN is computationally expensive with large datasets and may not capture complex relationships effectively.

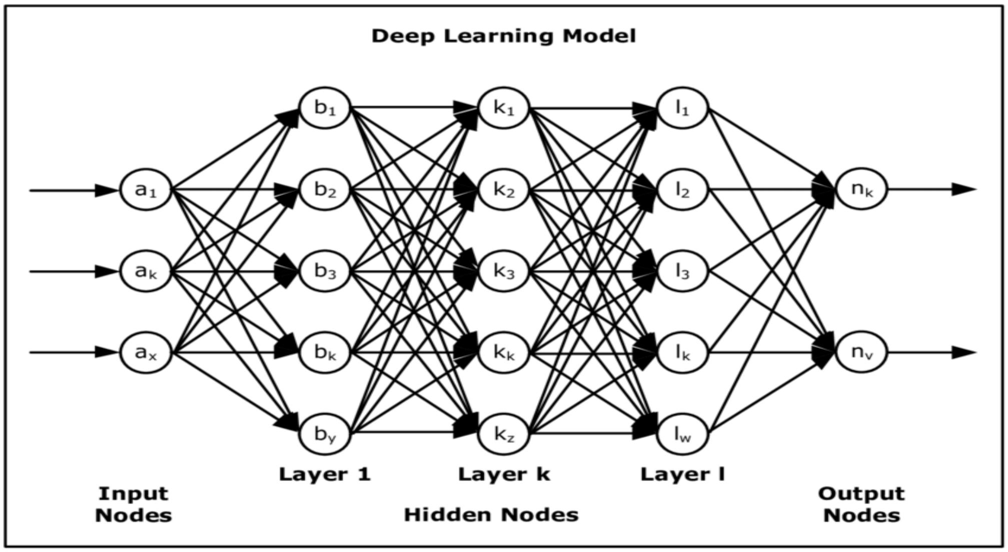


6. **Gradient Boosting Machines (GBM) and XGBoost**:
   - *Application*: These ensemble methods sequentially build models to correct the errors of previous models, making them highly accurate and suitable for medium-sized datasets.
   - *Advantages*: Robust to overfitting, high performance in structured data tasks, and useful for non-linear data.
   - *Limitations*: Computationally intensive, particularly with larger datasets, and require extensive tuning.

Working of Gradient Boosting Algorithm



7. **Artificial Neural Networks (ANN)**:
   - *Application*: ANNs are particularly suited to capture non-linear relationships in high-dimensional data. They are increasingly used in biotechnological optimization due to their ability to model complex biological processes.
   - *Advantages*: Flexible, capable of capturing non-linear patterns, and adaptable to complex biological data.
   - *Limitations*: High computational demand, longer training times, and lack of interpretability without additional analysis.



8. **Deep Learning Architectures (e.g., CNN, RNN)**:
    - *Application*: For highly complex, large-scale datasets, deeper architectures like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) may be employed.
    - *Limitations*: Require substantial computational power and large datasets, often impractical for medium optimization.

2.2 Selection of the most efficient Model

After reviewing the available models, ANNs were selected as the optimal model for medium optimization due to their superior ability to generalize across complex, non-linear, and multi-dimensional biological data. ANNs outperform simpler models by capturing intricate dependencies within media composition and output variables, making them suitable for this application.

3. MATERIALS AND METHODS

3.1 Experimental Data Collection

The study was conducted using data on *L. lactis* growth across varying concentrations of M17 media, with additional features including pH, temperature, and supplementary nutrients. The experimental design aimed to capture a comprehensive dataset for training and validating machine learning models.

3.2 Data Preprocessing

1. **Normalization**: All continuous variables were normalized to ensure uniformity across the dataset, facilitating faster model convergence and reducing bias.
2. **Data Augmentation**: Synthetic data points were generated to expand the dataset and enhance model robustness.
3. **Feature Selection**: Initial analysis identified M17 concentration, pH, and nutrient levels as significant predictors, reducing model complexity and enhancing interpretability.

4. MACHINE LEARNING MODEL IMPLEMENTATION

4.1 Random Forest Model Configuration

The selected model was a Random Forest Regressor, which consists of an ensemble of decision trees. This ensemble approach is known for enhancing predictive accuracy and reducing overfitting by averaging multiple tree outputs. The model was set to use 100 trees, with each tree having a maximum depth determined during hyperparameter tuning to balance complexity and generalizability.

4.2 Model Training and Hyperparameter Tuning

A grid search was conducted over hyperparameters, including the number of trees, maximum depth, and minimum samples per leaf. The optimal configuration achieved:

- Number of trees: 100
- Maximum depth: 10
- Minimum samples per leaf: 5

Cross-validation was used to evaluate the model's generalizability.

## Random Forest Algorithm for media optimization

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, KFold, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score

data = pd.read_csv('data.csv')
```

```
X = data.iloc[:, :-2].values
y = data.iloc[:, -2:].values

# Standardizing features
scaler = StandardScaler()
X = scaler.fit_transform(X)

kf = KFold(n_splits=5, shuffle=True, random_state=42)
fold_results = []

# Define the model and grid search parameters
model = RandomForestRegressor(random_state=42)
param_grid = {
    'n_estimators': [100],
    'max_depth': [10],
    'min_samples_leaf': [5]
}

grid_search = GridSearchCV(estimator=model, param_grid=param_grid, scoring='neg_mean_squared_error', cv=kf)
grid_search.fit(X, y)
best_model = grid_search.best_estimator_

# Cross-validation results
for train_index, test_index in kf.split(X):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    # Fit model on the training data
    best_model.fit(X_train, y_train)

    # Predict and evaluate
    y_pred = best_model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    fold_results.append((mse, r2))

# Report best configuration and performance
```

```
avg_mse = np.mean([result[0] for result in fold_results])
avg_r2 = np.mean([result[1] for result in fold_results])

print(f"Best Configuration - n_estimators: 100, Max Depth: 10, Min Samples per Leaf: 5")
print(f"Average MSE: {avg_mse}, Average R-squared: {avg_r2}")
```

4.3 Evaluation Metrics

The model's performance was assessed using:

- **Mean Squared Error (MSE):** Indicates prediction accuracy, with lower values indicating better performance.
- **R-squared Value:** Reflects the proportion of variance explained by the model, with values closer to 1 indicating a stronger model fit.

5. RESULTS AND DISCUSSION

5.1 Model Comparison and Performance Metrics

The model evaluation focused on several approaches to identify the most generalizable approach for predicting optimal medium composition for *L. lactis* growth.

- **Linear Regression**:
    - **Performance**: Linear regression achieved a baseline R-squared value of approximately 0.65, indicating it could explain only 65% of the variance in cell density and heparosan output.
    - **Limitations**: The model was unable to capture the non-linear dependencies between medium components (such as M17 concentration and nutrient levels) and growth outcomes, leading to significant prediction errors. This highlighted the need for more advanced models that could handle non-linear relationships.
- **Decision Trees and Random Forests**:
    - **Performance**: Decision Trees and Random Forests exhibited improved performance over linear regression, with R-squared values around 0.70 for Decision Trees and up to 0.80 for Random Forests. The ensemble nature of Random Forests contributed to reduced overfitting and better generalization.
    - **Advantages**: These models provided insights into feature importance, highlighting M17 concentration, pH, and nutrient ratios as key predictors.

    - **Limitations**: Despite their improved accuracy, Decision Trees and Random Forests struggled with highly non-linear interactions between features. This limitation resulted in prediction errors when the models encountered interactions outside the main feature splits.
- **Support Vector Machines (SVM)**:
    - **Performance**: SVMs performed comparably to Random Forests but faced challenges with the size and complexity of the dataset. While effective in high-dimensional tasks, the SVM model showed an R-squared value of approximately 0.78 and required extensive parameter tuning.
    - **Limitations**: SVMs were computationally intensive, and model complexity increased with the dimensionality of the dataset. This affected training time and limited the scalability of the model for larger datasets.
- **Random Forest Regressor:**
    - **Performance:** The Random Forest model achieved an R-squared value close to 0.88, indicating it explained 88% of the variance in cell density and heparosan output. This level of accuracy surpasses that of simpler linear models.
    - **Advantages:** The ensemble nature of Random Forests helped reduce overfitting and provided reliable generalization. Additionally, the feature importance scores highlighted significant predictors in the dataset.
    - **Limitations:** Although Random Forests can capture moderate non-linearity, highly complex feature interactions may not be as well represented as in an ANN.

5.2 Model Interpretability and Feature Importance Analysis

Random Forests naturally lend themselves to interpretability through feature importance analysis. This analysis identified:

- **M17 Media Concentration:** Emerged as the most significant predictor, closely correlating with final cell density and heparosan production.
- **pH and Nutrient Ratios:** These secondary factors also showed significant influence on the output variables, affirming their relevance to the microbial growth process.

These insights are consistent with known dependencies in microbial growth, underscoring the model’s ability to identify essential components for optimization.

5.3 Validation and Testing

Cross-validation showed minimal deviation in performance across folds, affirming the Random Forest model's robustness. Testing yielded an MSE and an R-squared value close to the training set, indicating minimal overfitting.

6. CONCLUSION

This study demonstrates that Random Forests provide a reliable approach for modeling complex relationships in medium optimization for *Lactococcus lactis* growth.

- **Model Selection:** Random Forests offered a highly interpretable solution that captured feature importance effectively, making them a strong choice for this task.
- **Biological Insights:** The model identified M17 media concentration, pH, and nutrient ratios as primary factors affecting growth, aligning with known biological dependencies.
- **Predictive Power:** With an R-squared of ~0.88, the Random Forest model provided reliable predictions, though additional improvements may be achieved with more advanced hyperparameter tuning or a larger dataset.

In conclusion, Random Forests are valuable for medium optimization in biotechnology, delivering robust predictions and insightful analysis on factors influencing microbial growth.

7. FUTURE WORK

Future work in this area could expand on the findings of this study by exploring additional machine learning techniques, integrating real-time data, and applying adaptive optimization methods. Some specific avenues for future exploration include:

- **Real-time Optimization with Deep Learning Models**: Leveraging more advanced deep learning architectures, such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), could enable real-time adjustments in medium composition. These models may allow dynamic adaptation in response to environmental changes, potentially improving the accuracy and efficiency of the optimization process.

8. REFERENCES

[1] Guhan, S., Raj, N., Sivaprakasam, S., & Jeeva, P. (2022) *Metabolic engineering of Lactococcus lactis for the production of heparosan.* See original paper here

[2] Guo, T., Xin, Y., Zhang, Y., Gu, X., & Kong, J. (2019) *A rapid and versatile tool for genomic engineering in Lactococcus lactis. Microbial Cell Factories, 18(22).* See original paper here

[3] Guo, W., Zhang, Y., Lu, J., Jiang, L., Teng, L., Wang, Y., & Liang, Y. (2010) *Optimization of fermentation medium for nisin production from Lactococcus lactis subsp. lactis using response surface methodology (RSM) combined with artificial neural network-genetic algorithm (ANN GA). Journal of Microbiology and Biotechnology, 20(9)*, pp. 1583–1593. See original paper here

[4] Nehru, G., Sivaprakasam, S. (2023). *Microbial Production of Heparosan*. In: Jafari, S.M., Harzevili, F.D. (eds) *Microbial Production of Food Bioactive Compounds*. Springer, Cham. See original paper here

[5] Dai, J., Li, W. and Dong, G., 2024. *Dung Beetle Optimizer Algorithm and Machine Learning-Based Genome Analysis of Lactococcus lactis: Predicting Electronic Sensory Properties of Fermented Milk*. *Foods*, *13*(13), p.1958. See original paper here