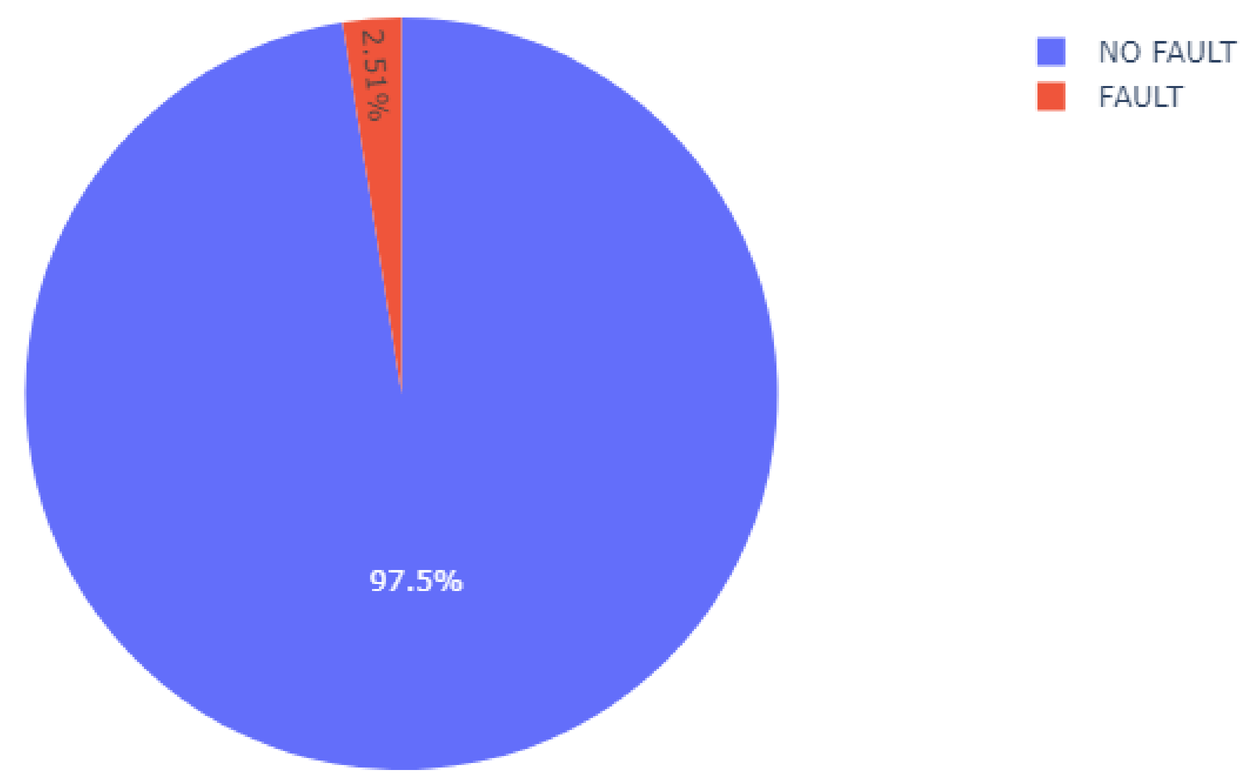
Pharma Production AI-ML Case Study

This data was generated using an advanced mathematical simulation of a 100,000 litre penicillin fermentation system referenced as IndPenSim (Industrial Penicillin Simulation).

IndPenSim is the first simulation to include a realistic simulated Raman spectroscopy device for the purpose of developing, evaluating and implementation of advanced and innovative control solutions applicable to biotechnology facilities. This data set generated by IndPenSim represents the biggest data set available for advanced data analytics and contains 100 batches with all available process and Raman spectroscopy measurements (~2.5 GB).

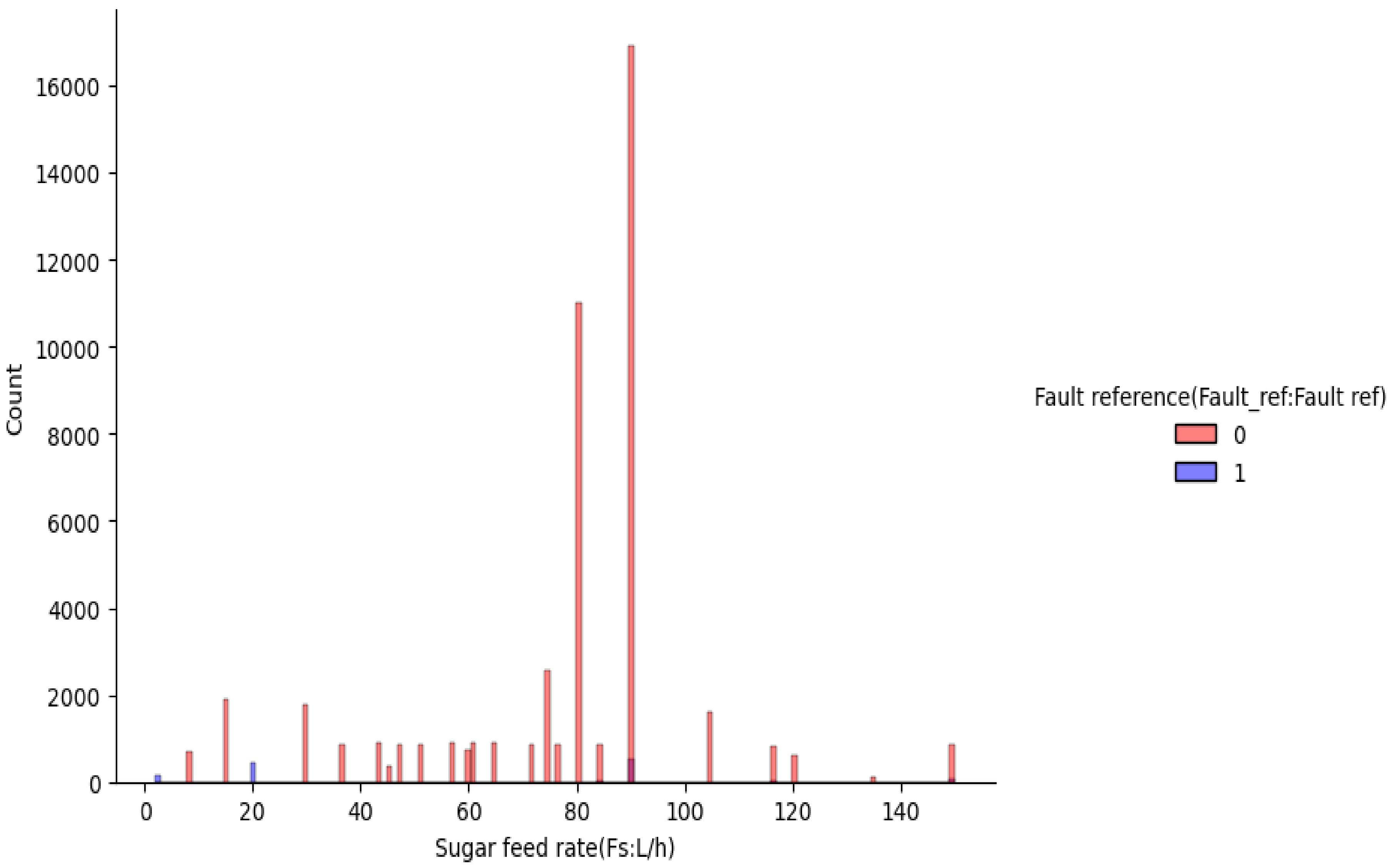
This data is highly suitable for the development of big data analytics, machine learning (ML) or artificial intelligence (AI) algorithms applicable to the biopharmaceutical industry. The 100 batches are controlled using different control strategies and different batch lengths representing a typical Biopharmaceutical manufacturing facility.

The main goal is to predict the faulty reference and the factors that determine the Fault Reference using Auto-ML Model.

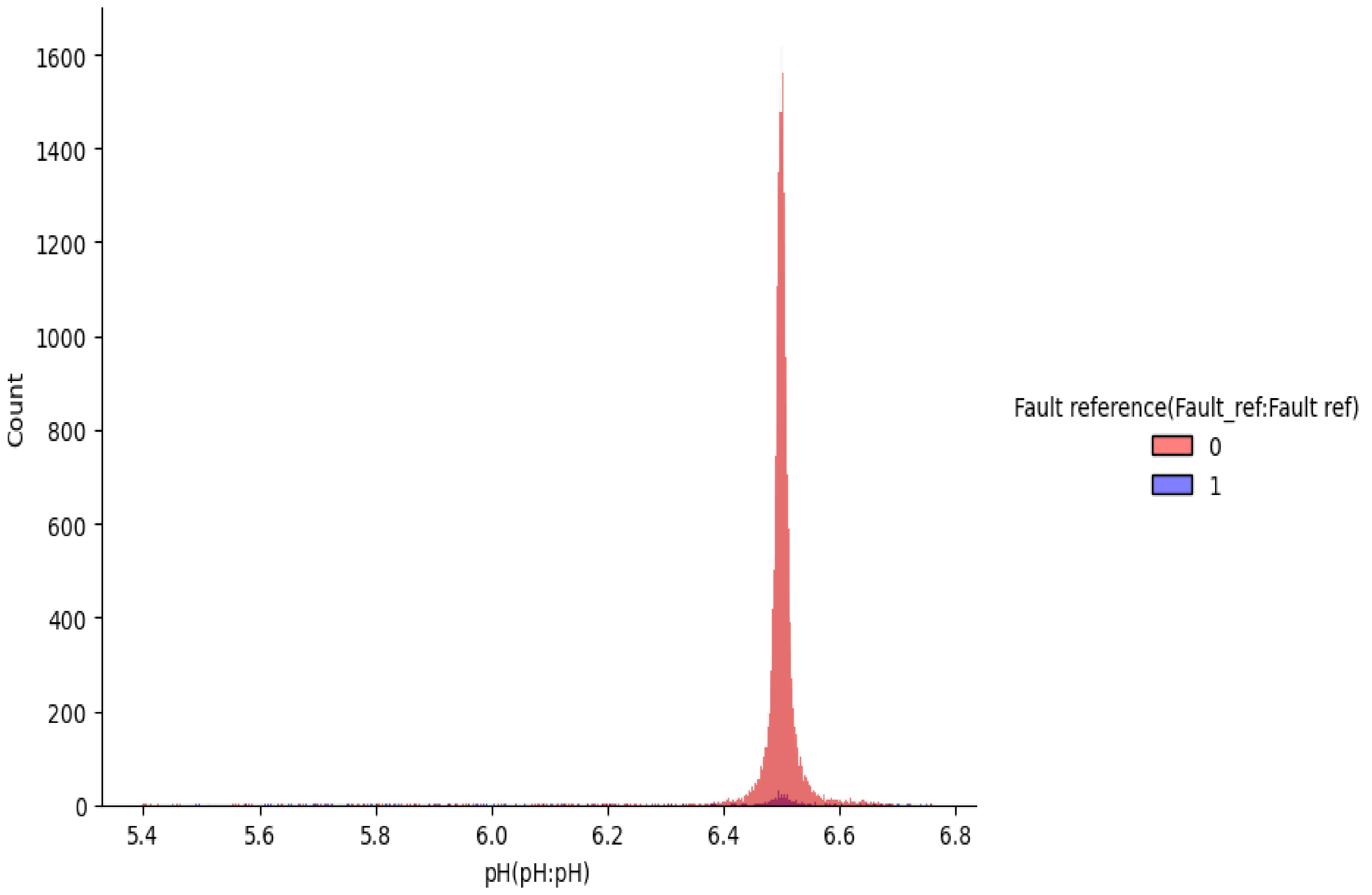

Class Distribution

Fault Reference	No. of Candidates
No Fault	48744
Fault	1256
No Fault Rate	97.5 %
Fault Rate	2.51 %

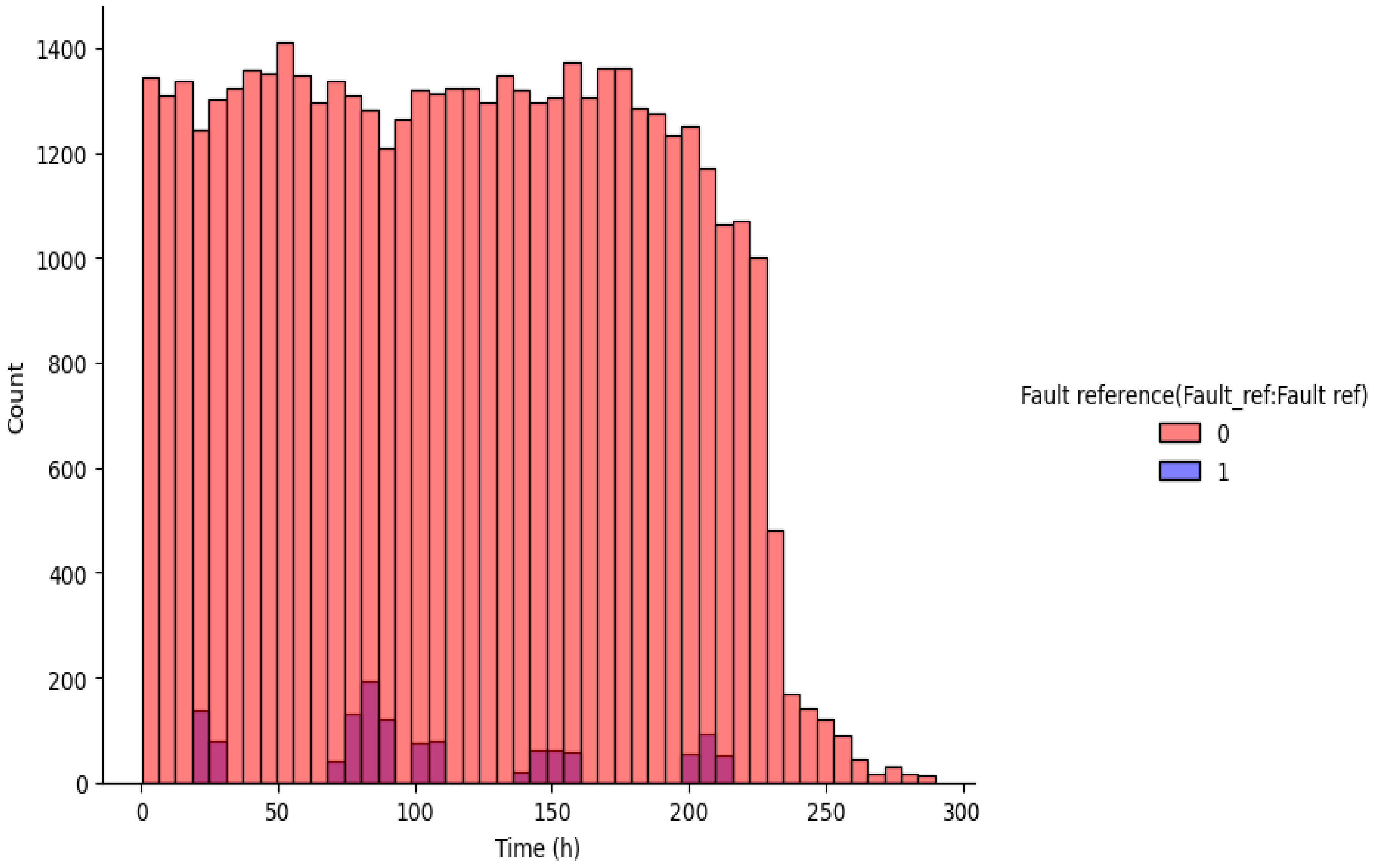
Features Responsible

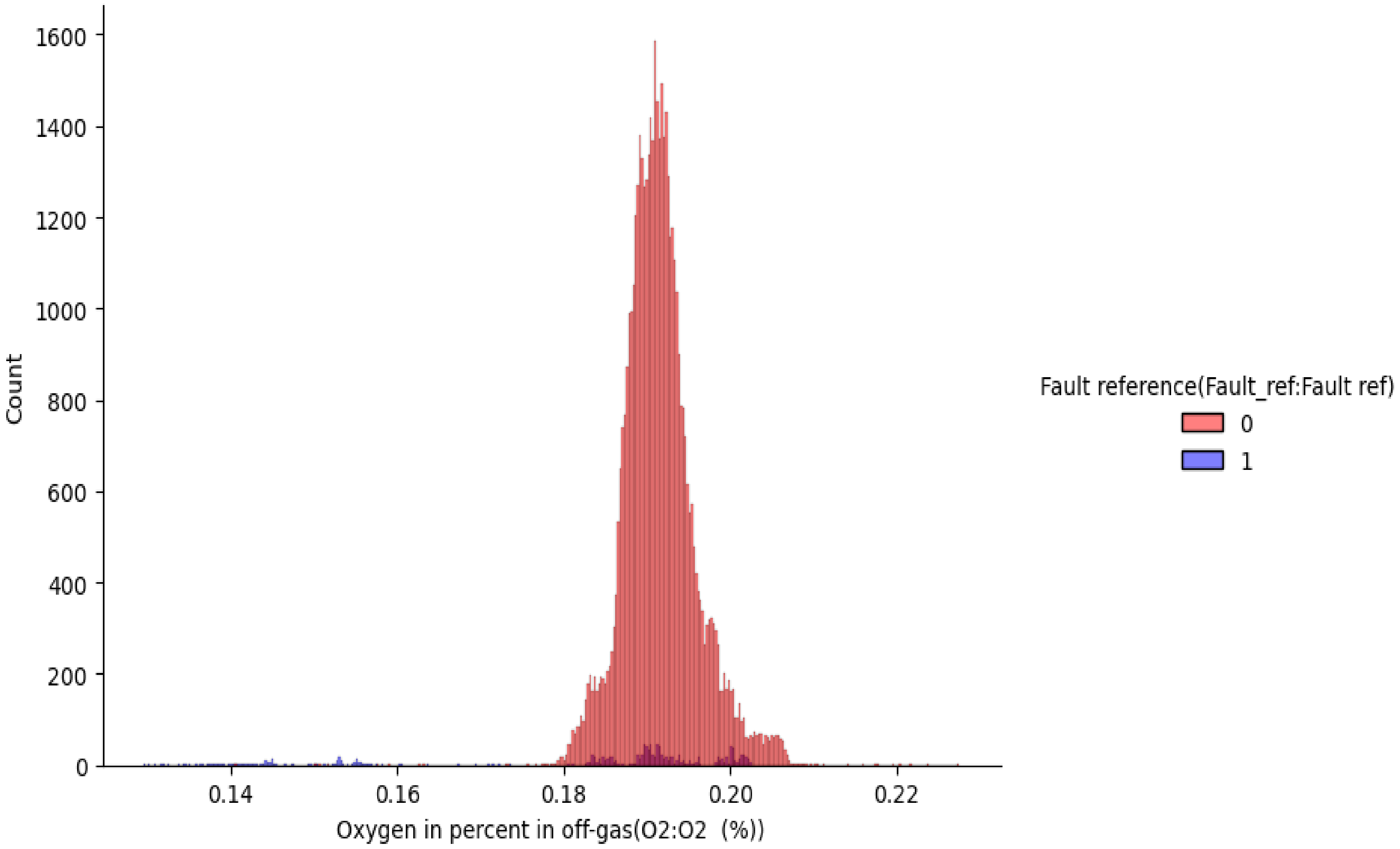


- **Sugar feed rate(Fs:L/h)**
- The sugar feed rate impacts the growth and metabolism of the Penicillium chrysogenum mold. If the sugar feed rate is too low, the mold may not have enough carbon to grow and produce penicillin at the desired rate. On the other hand, if the sugar feed rate is too high, it can lead to an overgrowth of the mold and the production of undesirable byproducts, which can reduce the yield and quality of the penicillin.



- **pH(pH:pH)**
- The pH level impacts the growth and metabolism of the Penicillium chrysogenum mold. The optimal pH range for penicillin production is typically between pH 6.0 and 8.0. If the pH level is too low or too high, it can impact the growth of the mold, as well as the activity of enzymes involved in the penicillin biosynthesis pathway.





- **Time (h)**
- Time is a critical factor that can impact the industrial-scale production of penicillin. Optimizing the duration of the fermentation process is essential to ensure the maximum production of penicillin with the highest possible yield and quality

- **Oxygen in percent in off-gas(O2:O2 (%))**
- If the percentage of oxygen in the off-gas is too low, it can limit the growth of the mold and reduce the yield and quality of the penicillin. On the other hand, if the percentage of oxygen in the off-gas is too high, it can lead to the overproduction of reactive oxygen species, which can damage the mold cells and reduce the yield and quality of the penicillin.

Auto-ML Methodology Results

Algorithms	Test Accuracy (25 percentile)	Test Accuracy (50 percentile)	Test Accuracy (75 percentile)	Test Accuracy (90 percentile)
Decision Tree	100	99.59	99.94	98.14
Random Forest	100	100	100	100
XGBoost	100	100	100	100
MLP	50.96	79.50	60.76	64.62
RNN	50	50	50	50
Total Features	25	49	74	89
Avg. Accuracy	80.192	85.818	82.14	82.552

- **Based on our observation from the standard ML algorithms, 50 percentile has the best average accuracy**
- **XGBoost and Random Forest was the best performing algorithm across all percentile with 100% accuracy.**

Conclusion

In Pharma Production Industry, Auto-ML can help in the process of drug development and production by making it more effective. Auto-ML can provide diagnostic assistance and empower physicians to deliver personalized treatment. The dataset contains 113,935 records with 1 Categorical feature and 99 Numerical features.

For classification, models were created with algorithms using Auto-ML techniques like Decision tree, Multilayer Perceptron, Random forest and XGBoost . With these models, performance measurement values were obtained for feature sets of 25, 49, 74 and 89. The Auto-ML algorithms were able to predict the Fault Reference based on their features with an average accuracy between 80% – 83% and helped to identify factors that determine the Fault Reference.

The major factors include Sugar Feed Rate, pH, Time and Oxygen percentage in off gas. When the results are examined, it is observed that with the addition of each new feature, the success of classification decreases. Based on the performance measurement values obtained, it is possible to say that the study achieved success in classifying fault reference .