

TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Omar Saïd Mohamed El Nahhas 221136IAFM

**EXPLAINABLE AI-BASED RECOMMENDATION MODEL
FOR YIELD IMPROVEMENT OF A PHARMACEUTICAL API
SYNTHESIS PROCESS**

Master Thesis

Academic supervisor

Sven Nõmm, PhD

Company supervisor

Dr.-Ing. Christoph Bergs

Tallinn 2022

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Omar Saïd Mohamed El Nahhas 221136IAFM

**XAI PÕHINEV SOOVITUSMUDEL FARMATSEUTILISE API
SÜNTEESIPROTSESSI SAAGISE SUURENDAMISEKS**

Magistritöö

Akadeemiline juhendaja

Sven Nõmm, PhD

Ettevõtte juhendaja

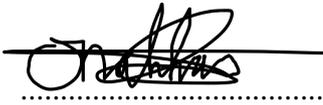
Dr.-Ing. Christoph Bergs

Tallinn 2022

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Omar Saïd Mohamed El Nahhas

A handwritten signature in black ink, appearing to read 'Omar Saïd Mohamed El Nahhas', is written over a solid horizontal line. Below this line is a dotted horizontal line.

(signature)

Date: May 9, 2022

Extended summary

This research aimed to determine if a prediction model can provide feature value recommendations to improve the yield of an active pharmaceutical ingredient synthesis process while maintaining purity. The black-box production process made yield improvement efforts too complex and expensive, which caused an inconsistent yield output, leading to a significant financial loss for every unachieved yield percentage. Current literature on yield optimisation showed a strong presence in agricultural settings, with a significantly lower count in manufacturing, and yield optimisation or improvement research focussed on active pharmaceutical ingredient synthesis in pharmaceutical manufacturing being extremely scarce. Moreover, current available literature in yield optimisation in other fields were stuck in the dilemma of explainability versus performance of the model, causing the most accurate results to remain a black-box, and the most transparent models to perform sub-par. Meanwhile, observed research on active pharmaceutical ingredient synthesis yield improvement had not actively adopted machine learning methods yet.

This thesis focussed on the application of machine learning methods and algorithms to active pharmaceutical ingredient synthesis yield improvement, which were previously not used due to the intransparency of the models coupled with strict drug manufacturing regulations. Therefore, the novelty of this work was the application and comparison of machine learning and explainable artificial intelligence methods to approximate the inner-workings of the black-box active pharmaceutical ingredient synthesis process for yield improvement, closing the knowledge gap related to the dilemma of explainability versus performance of the model, as well as introducing machine learning methods from related fields to active pharmaceutical ingredient synthesis yield improvement. The novelty of this work is visualised in Fig. 1, with unrelated fields (according to the observed literature) depicted as having no overlap.

Summarizing, this thesis focussed on the large gap between active pharmaceutical ingredient synthesis yield improvement and machine learning and explainable artificial intelligence. This project was structured according to the cross industry standard process for data mining methodology.

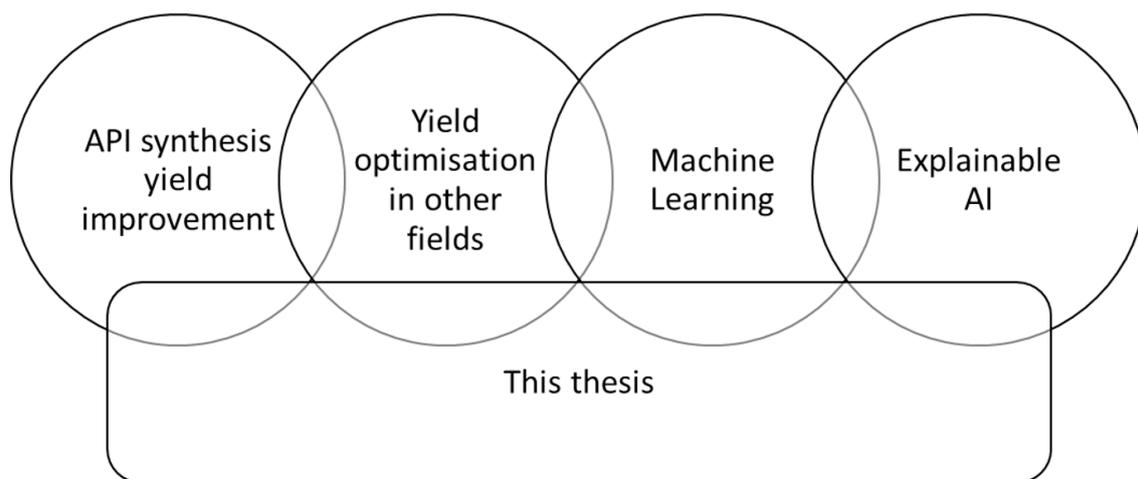


Figure 1. Overview of the proposed novelty of this thesis

Data from programmable logic controllers and chemists' archives were manually extracted. Then, the heterogeneous data sources were analysed using descriptive analytics, consisting of numerical summaries, distribution plots, correlation plots and time-series plots. To prepare the data for modelling, the separate data sources were merged into a single homogeneous dataset using data type porting, where the mismatch of sampling frequency between sensor data and production data was approached with simple feature extraction methods such as the mean, maximum and standard deviation to match the slowest sampling frequency across all data sources. Afterwards, feature selection was performed using process expertise, enabling the elimination of variables with no significance for this research, as well as determining which features were physically controllable and which were not. All collected data led to batches with an acceptable purity. All batches with a significant amount of missing data were removed. Randomly missing values due to measurement failures were replaced with zero rather than an imputed value, keeping all data true-to-life to ensure an acceptable purity of the yield.

To approximate the black-box of the underlying industrial process, yield models were created using linear and non-linear machine learning models and explainable artificial intelligence methods to uncover the importance and contribution of the top 15 influential process features on the yield, which led to six directly applicable yield improvement recommendations. For modelling the yield, Elastic Net Regularisation, Random Forest and a Voting Ensemble model found through automated machine learning (utilising XGBoost, LightGBM and Random Forest models internally) were used. Using all features, the Voting Ensemble model performed best at predicting the yield with an R^2 of 88%, root mean squared error of 0.39 and mean absolute error of 0.21. Using the Elastic Net and Random Forest embedded capabilities, and the Voting Ensemble with explainable artificial intelligence methods surrogate, permutation, Shapley additive explanations and

local interpretable model-agnostic explanations, the top 15 most important features were determined through a feature sensitivity analysis. A yield model was then trained with the top 15 features, resulting in a Voting Ensemble model predicting the yield with an R^2 of 91%, root mean squared error of 0.30 and mean absolute error of 0.16, which indicated a successful approximation of the underlying black-box of the production process.

To make the recommendations personalised to the raw material quality and desired yield as input, a model was built which predicted the values of the top 15 features in order to obtain the desired yield. The used methods were Deep Neural Network, AdaBoost, Elastic Net Regularisation, Kriging, Support Vector Regression, LightGBM, Gradient Boosting, Voting Ensemble (with Kriging, Random Forest and Gradient Boosting), Bagging, XGBoost, Decision Tree and Extra Trees, with the Extra Trees performing the best with an R^2 of 79% and normalised average root mean squared error of 0.07 across all top 15 features. Consequently, with the top 15 features determined by a model with an R^2 of 91%, the personalised recommendations to improve the yield explained 72% of the data's variance and had an average error margin of 7% per feature. Concluding, the recommendation model is capable of providing personalised feature values for yield improvement while maintaining purity, which explained a total of 72% of the black-box of the production process and its control.

This thesis closed the knowledge gap of machine learning and explainable artificial intelligence applied to active pharmaceutical ingredient synthesis yield improvement, and provided the developed tools and theoretical fundamentals in order to increase the mean and reduce the standard deviation of the active pharmaceutical ingredient synthesis yield output, leading to an immediate yearly increase in profit of millions of euros when applied. An interesting topic for further research is the question if the production data from years ago is still relevant today due to concept drift, and if it is useable for future predictions. Due to this uncertainty, the decision was made to take the latest months of data and work with a smaller data set instead, making some model architectures such as a Deep Neural Network more prone to over-fitting. Moreover, future research should test more elaborated time-series feature extraction methods and compare performance with the method in this work to match sampling frequencies. Finally, approaching this problem from a control theory perspective was not considered in this work. Mechanistic modelling of the process is complex, time-consuming, and requires expertise knowledge in chemistry, control theory and the underlying process. However, mechanistic models are potentially more accurate than data-driven models, as they fully capture the underlying physical properties of the production process when executed correctly. Consequently, future research should focus on hybrid modelling, combining data-driven modelling as presented in this approach, with mechanistic model components to obtain the best performing model.