Heterogeneous Transfer Learning from Batch to Continuous Direct Compression Tablet Manufacturing

Yuki Kobayashi [*] **Takuya Nagato** [**] **Takuya Oishi** [**,****]
Sanghong Kim [***] **Shota Kato** [*] **Manabu Kano** [*]

[*] *Department of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: manabu@human.sys.i.kyoto-u.ac.jp)*
[**] *Powrex Corporation, Hyogo 664-0837, Japan*
[***] *Department of Applied Physics and Chemical Engineering, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan*
[****] *Department of Applied Chemistry, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan*

Abstract: There is growing interest in shifting from batch manufacturing to continuous manufacturing in the pharmaceutical industry. Constructing statistical models that accurately predict critical quality attributes (CQAs) from operating conditions with minimal experiments in the new process is required to identify the optimal operating conditions and monitor the process. This study aims to demonstrate that heterogeneous transfer learning (TL) using data from the batch direct compression (BDC) process can enhance the prediction performance of CQAs in the continuous direct compression (CDC) process. We conducted 26 BDC experiments and 19 CDC experiments. Predictive models of tablet hardness were then built using partial least squares regression, Gaussian process regression, and random forest regression. We employed frustratingly easy heterogeneous domain adaptation (FEHDA) to the two experimental datasets, treating BDC as the source domain and CDC as the target domain. We found that FEHDA achieved lower RMSE and higher R^2 than those by models trained using only the CDC dataset. RFR attained the best predictive performance with an average RMSE improvement of 9.36 N. Notably, FEHDA improved the prediction performance in the region where no samples were obtained from the BDC process. These results support the effectiveness of heterogeneous TL for the shift from BDC to CDC.

Keywords: Process modeling, Transfer learning, Domain adaptation, Pharmaceutical manufacturing, Continuous manufacturing, Solid dosage, Direct compression

1. INTRODUCTION

In the pharmaceutical industry, a shift from batch manufacturing (BM) to continuous manufacturing (CM) has been gaining attention (Food and Drug Administration, 2019) because CM offers several advantages, including adaptability to fluctuations in supply demands and a smaller footprint. A successful transition from BM to CM requires determining the range of operating conditions of the new CM process that achieves critical quality attributes (CQAs) of the product within the specifications. One of the effective approaches is to construct statistical models representing the relationship between operating conditions and CQAs. The accuracy of statistical models tends to improve as the amount of data increases; however, conducting a sufficient number of experiments is difficult due to the high cost of raw materials, such as active pharmaceutical ingredients (APIs). Thus, it is essential to develop strategies for building models with high predictive performance using limited data available in new CM processes.

Transfer learning (TL) (Pan and Yang, 2010) can be a solution to this data shortage problem. Many machine learning methods assume that the training and test data are drawn from the same feature space and distribution. TL is a technique that improves the performance of models in a target domain (TD) with a small amount of data by transferring knowledge from a source domain (SD), where data are either already available or inexpensive to obtain. When the feature spaces differ between the SD and TD, TL methods are specifically referred to as heterogeneous TL.

Heterogeneous TL approaches are suitable for the shift from BM to CM since BM and CM have different process variables. Although BM and CM differ in their equipment and operating methods, they have similarities in the mechanisms and objectives of unit operations, suggesting the potential for transferring knowledge from BM to CM.

TL has been applied to manufacturing processes (Bang et al., 2019). In the context of tablet manufacturing processes, TL was used for the scale-up of tablet presses (Yaginuma et al., 2024). However, there is no research on

applying TL to the transition from BM to CM nor to direct compression processes, which are one of the primary methods of tablet manufacturing.

This study aims to demonstrate the effectiveness of TL using data on the batch direct compression (BDC) process in improving the prediction performance of CQAs in the continuous direct compression (CDC) process. We first conducted two sets of experiments: one using the BDC process and the other using the CDC process. Then, we employed frustratingly easy heterogeneous domain adaptation (FEHDA) (Kobayashi et al., 2022), a method of the heterogeneous TL, to build statistical models for predicting the tablet hardness as a CQA. The datasets collected in the BDC and CDC process experiments were used as the SD and TD, respectively. Finally, we compared the prediction performance of these models with that of models constructed using only the TD dataset.

2. MATERIALS AND METHODS

2.1 Experiments

To collect data from the two different domains, two sets of experiments were conducted using the CDC process and the BDC process. In all experiments, Acetaminophen (APAP) (Spera Nexus, Japan) was used as an API, SuperTab 11SD (DFE Pharma, Germany) was used as an excipient, and magnesium stearate (MgSt) (Taihei Chemical Industrial, Japan) was used as a lubricant.

In the CDC process, the API and excipient were fed into the continuous API mixer (MG100, Powrex, Japan) by two screw feeders (LIW-300-P, Ishida, Japan). MgSt and the intermediate product from the API mixer were fed into the continuous lubricant mixer (MG100, Powrex, Japan). The intermediate product from the lubricant mixer was fed into the rotary tablet press (FETTE 102i, Fette Compacting, Germany) to produce tablets.

In the BDC process, the API and excipient were weighed and then fed into the mini-batch API mixer (MG200, Powrex, Japan). After mixing for a specified time, weighed MgSt were added to the MG200, and it was used as the batch lubricant mixer. After mixing for another specified time, the intermediate product from the lubricant mixer was fed into the rotary tablet press (FETTE 102i, Fette Compacting, Germany) to produce tablets.

The CDC process has nine process variables; five variables are common to CDC and BDC, and the other four variables are CDC-specific, as shown in Table 1. Similarly, the BDC process has 11 process variables; five variables are common to CDC and BDC, and the other six variables are BDC-specific. The variables were set at three levels in the ranges presented in Table 1, and the combinations of their values were determined based on a definitive screening design (Jones and Nachtsheim, 2011). The design for nine variables was utilized for the CDC process, and the design for twelve variables with a center run was used for the BDC process. As a result, 19 samples for the CDC process and 26 samples for the BDC process were obtained.

In each experiment, the tablet hardness [N] was measured as a CQA with the hardness measurement machine (TBH 425 TD, Erweka, Germany).

2.2 CQA prediction models

BDC was used as an SD, and CDC was used as a TD. The process variables in Table 1 were used as input variables of models. In the following, N_s is the number of samples in the SD, N_t is the number of samples in the TD, p_s is the number of variables in the SD, p_t is the number of variables in the TD, and p_c is the number of variables common to both domains.

As shown in Table 1, the input variables differ in the CDC and BDC processes. Thus, we applied FEHDA as a TL method. In FEHDA, the input variable matrix of the SD is divided into $X_\mathrm{c}^{(\mathrm{s})} \in \mathbb{R}^{N_\mathrm{s} \times p_\mathrm{c}}$ and $X_\mathrm{u}^{(\mathrm{s})} \in \mathbb{R}^{N_\mathrm{s} \times (p_\mathrm{s} - p_\mathrm{c})}$, which are the matrices of the variables common to both domains and SD-specific variables, respectively. Similarly, the input variable matrix of the TD is divided into $X_\mathrm{c}^{(\mathrm{t})} \in \mathbb{R}^{N_\mathrm{t} \times p_\mathrm{c}}$ and $X_\mathrm{u}^{(\mathrm{t})} \in \mathbb{R}^{N_\mathrm{t} \times (p_\mathrm{t} - p_\mathrm{c})}$, which are the matrices of the variables common to both domains and TD-specific variables, respectively. FEHDA defines the matrix of input variables as follows:

$$X = \begin{pmatrix} X_\mathrm{c}^{(\mathrm{s})} & X_\mathrm{c}^{(\mathrm{s})} & X_\mathrm{u}^{(\mathrm{s})} & \mathrm{O} & \mathrm{O} \\ X_\mathrm{c}^{(\mathrm{t})} & \mathrm{O} & \mathrm{O} & X_\mathrm{c}^{(\mathrm{t})} & X_\mathrm{u}^{(\mathrm{t})} \end{pmatrix}, \tag{1}$$

where O is the zero matrix. The output variable vector $\mathbf{y}$ is expressed as follows:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(\mathrm{s})} \\ \mathbf{y}^{(\mathrm{t})} \end{pmatrix}, \tag{2}$$

where $\mathbf{y}^{(\mathrm{s})} \in \mathbb{R}^{N_\mathrm{s}}$ and $\mathbf{y}^{(\mathrm{t})} \in \mathbb{R}^{N_\mathrm{t}}$ are output variable vectors for the SD and TD, respectively.

The preprocessing was performed for both X and $\mathbf{y}$. In the preprocessing of X, columns containing data from both domains were standardized together. For other columns, $X_\mathrm{c}^{(\mathrm{s})}$, $X_\mathrm{u}^{(\mathrm{s})}$, $X_\mathrm{c}^{(\mathrm{t})}$, and $X_\mathrm{u}^{(\mathrm{t})}$ were first respectively standardized, and then zero matrices were added. In the preprocessing of $\mathbf{y}$, the data from both domains were standardized together.

To investigate how FEHDA works with different regression methods, we built CQA prediction models using three methods: partial least squares regression (PLSR) (Geladi and Kowalski, 1986), Gaussian process regression (GPR) (Rasmussen and Williams, 2005), and random forest regression (RFR) (Breiman, 2001). PLSR is a linear regression method using latent variables (LVs). The number of LVs is a hyperparameter. GPR is one of the kernel methods. We used a kernel composed of a constant kernel, radial basis function kernel, and white kernel, as follows:

$$k(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = c \exp\left(-\frac{d(\boldsymbol{x}_n, \boldsymbol{x}_{n'})^2}{2l^2}\right) + \sigma\delta_{nn'}, \tag{3}$$

where $\boldsymbol{x}_n$ is the input variable vector of the n^{th} sample, $d(\boldsymbol{x}_n, \boldsymbol{x}_{n'})$ is the Euclidean distance between $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$, and $\delta_{nn'}$ is a Kronecker delta. c, l, and σ are hyperparameters. RFR is one of the ensemble learning methods using multiple decision trees. The number of trees and the maximum depth of decision trees were chosen as hyperparameters.

The evaluation criteria of models are the root mean square error (RMSE) and the coefficient of determination (R^2).

Table 1. Process variables changed in the experiments using the CDC and BDC processes.

Variable Name	Common to CDC & BDC	CDC-specific	BDC-specific	Range
API content	✓	-	-	5–15%
Production speed	✓	-	-	15–25 kg/h
Force feeder speed	✓	-	-	10–50 rpm
Pre-compression to main-compression force ratio	✓	-	-	20–60%
Main-compression force	✓	-	-	5–20 kN
Continuous API mixer center blade speed	-	✓	-	500–2500 rpm
Continuous API mixer scraper blade speed	-	✓	-	30–70 rpm
Continuous lubricant mixer center blade speed	-	✓	-	100–1000 rpm
Continuous lubricant mixer scraper blade speed	-	✓	-	30–70 rpm
Batch API mixer center blade speed	-	-	✓	500–1500 rpm
Batch API mixer scraper blade speed	-	-	✓	20–80 rpm
Batch API mixing time	-	-	✓	30–90 s
Batch lubricant mixer center blade speed	-	-	✓	100–500 rpm
Batch lubricant mixer scraper blade speed	-	-	✓	20–50 rpm
Batch lubricant mixing time	-	-	✓	30–60 s

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}, \tag{4}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2}, \tag{5}$$

where N is the number of the samples, y_n is the n^{th} actual value, $\hat{y}_n$ is the n^{th} predicted value, and $\bar{y}$ is the mean of the output variable.

The procedure of model construction and evaluation is as follows:

1 Randomly split the TD data into a training dataset $D^{(\mathrm{t})}_{\mathrm{train}}$ with nine samples and a test dataset $D^{(\mathrm{t})}_{\mathrm{test}}$ with ten samples.
2 Prepare X and $\mathbf{y}$ of FEHDA using the SD data and $D^{(\mathrm{t})}_{\mathrm{train}}$.
3 Perform preprocessing of X and $\mathbf{y}$.
4 Select a regression method from PLSR, GPR, or RFR.
5 Build a model using the preprocessed X and $\mathbf{y}$.
6 Calculate RMSE and R^2 using $D^{(\mathrm{t})}_{\mathrm{test}}$.
7 Perform Steps 4 to 6 for PLSR, GPR, and RFR.
8 Perform Steps 1 to 7 ten times changing the samples in $D^{(\mathrm{t})}_{\mathrm{train}}$ and $D^{(\mathrm{t})}_{\mathrm{test}}$.

In Step 5, hyperparameters were determined by leave-one-out cross-validation (LOOCV) in PLSR and RFR and by maximum likelihood estimation in GPR.

To verify the benefits of TL, we also built models using only $D^{(\mathrm{t})}_{\mathrm{train}}$ and validated the prediction performance with $D^{(\mathrm{t})}_{\mathrm{test}}$. This method using only the TD data is referred to as only target (OT) in the following. The procedure for model construction of OT is the same as that of FEHDA except for Steps 2 and 3. In Step 2, the input variable matrix and the output variable vector are formed as $X = (X^{(\mathrm{t})}_{\mathrm{c}}, X^{(\mathrm{t})}_{\mathrm{u}})$ and $\mathbf{y} = \mathbf{y}^{(\mathrm{t})}$, respectively. In Step 3, X and $\mathbf{y}$ are standardized.

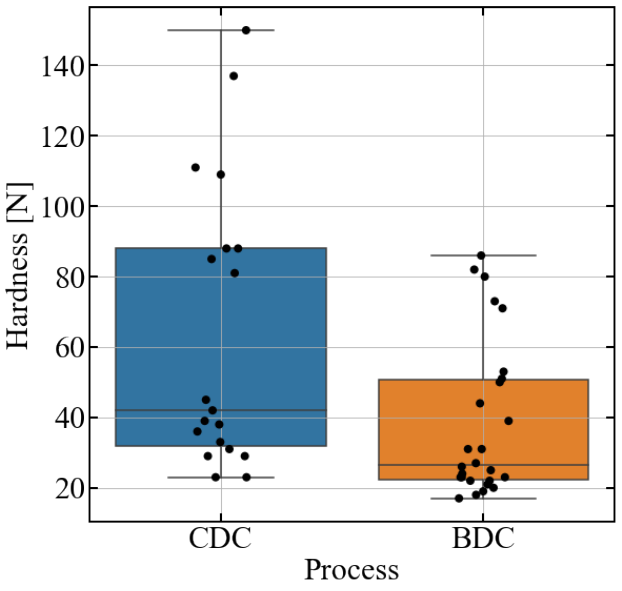

Fig. 1. Tablet hardness measured in CDC and BDC processes.

3. RESULTS AND DISCUSSIONS

3.1 Experiments

Figure 1 shows the hardness of tablets manufactured by the CDC and BDC processes. Since the formulation and the range of common variables are the same in both processes, the difference in the distributions of the hardness is expected to be explained by the process-specific variables, suggesting that heterogeneous TL should be applied.

3.2 CQA prediction models

Figure 2 presents the RMSE and R^2 of three regression methods using FEHDA and OT for $D^{(\mathrm{t})}_{\mathrm{test}}$ across ten splits. For all regression methods, FEHDA has lower averages of RMSE and higher averages of R^2 than OT. Specifically, the average RMSEs of FEHDA are lower than those of

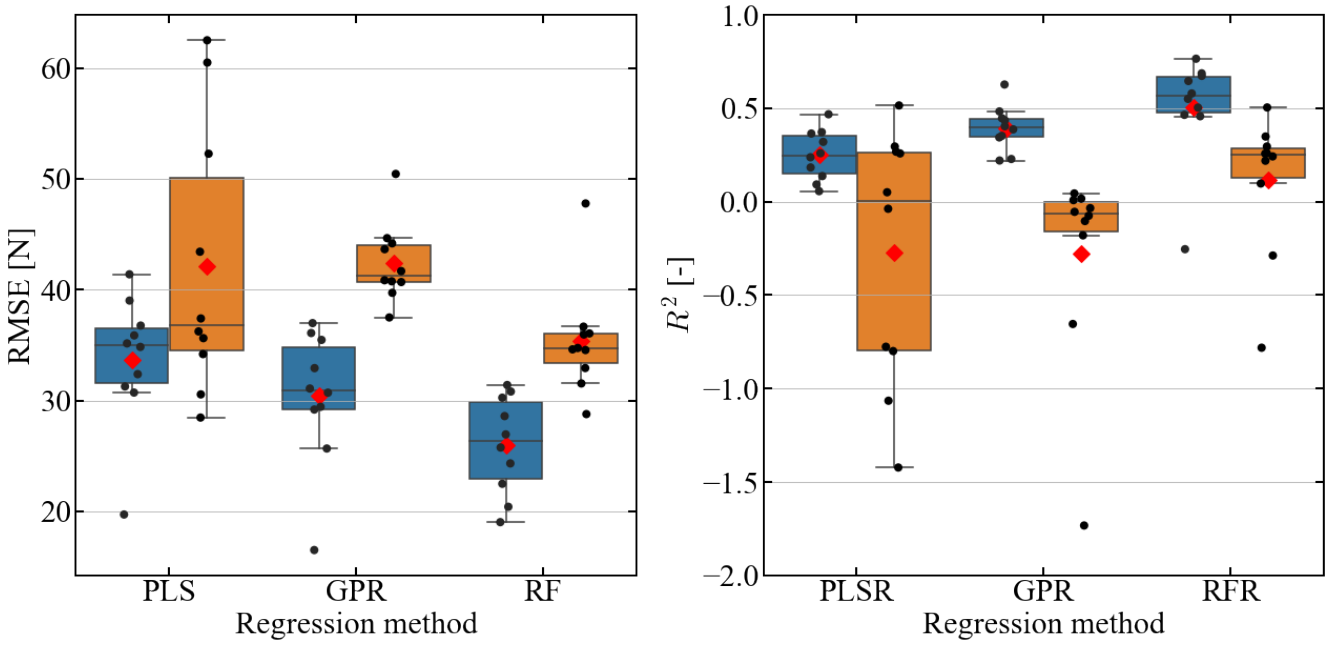

Fig. 2. RMSE (left) and R^2 (right) of three regression methods using FEHDA (blue) and OT (orange). The red diamond represents the average value.

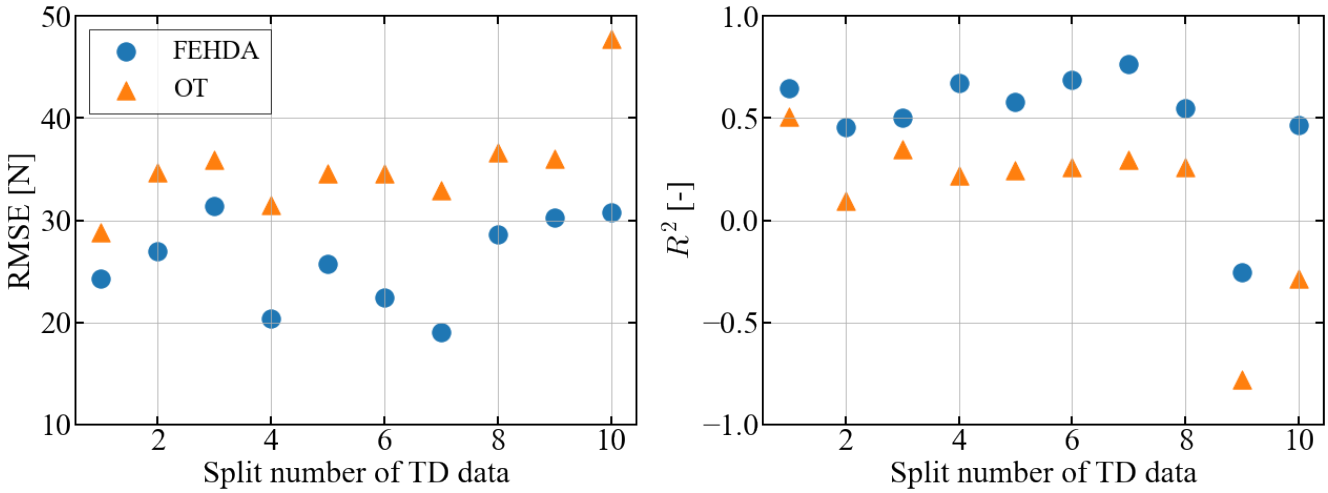

Fig. 3. RMSE (left) and R^2 (right) by RFR models built using FEHDA and OT across each of the ten splits of the TD data.

OT, with values of 8.40 N for PLS, 12.0 N for GPR, and 9.36 N for RFR. Among the regression methods using FEHDA, RFR has the lowest average RMSE and highest average R^2. Figure 3 illustrates RMSE and R^2 by RFR models using FEHDA and OT for $D_{\text{test}}^{(\text{t})}$ in each of the ten splits of the TD data. In all splits, FEHDA outperforms OT on both criteria. The following paragraph presents a discussion of the results obtained from Splits 7 and 9, in which FEHDA exhibits the highest and lowest R^2, respectively.

Figure 4 displays the actual and predicted hardness of $D_{\text{test}}^{(\text{t})}$ by the RFR models using FEHDA and OT for Splits 7 and 9. For most samples, the plots of FEHDA are closer to the diagonal line than those of OT. This means that FEHDA improves their prediction performance. Regarding the region where the hardness is above 90 N, extrapolation occurs in the sense that the maximum hardness in $D_{\text{test}}^{(\text{t})}$ is higher than that in $D_{\text{train}}^{(\text{t})}$ in Split 7. Despite the absence of samples with hardness above 90 N in the SD data as shown in Figure 1, FEHDA improves the prediction performance in the region. This suggests that the performance improvement is not due to increased sample density in the region with limited TD data. In the region where the hardness is below 35 N, Figure 5 shows that extrapolation also occurs as the minimum hardness in $D_{\text{test}}^{(\text{t})}$ is lower than that in $D_{\text{train}}^{(\text{t})}$ in Split 9. According to Figure 1, the SD data contain many

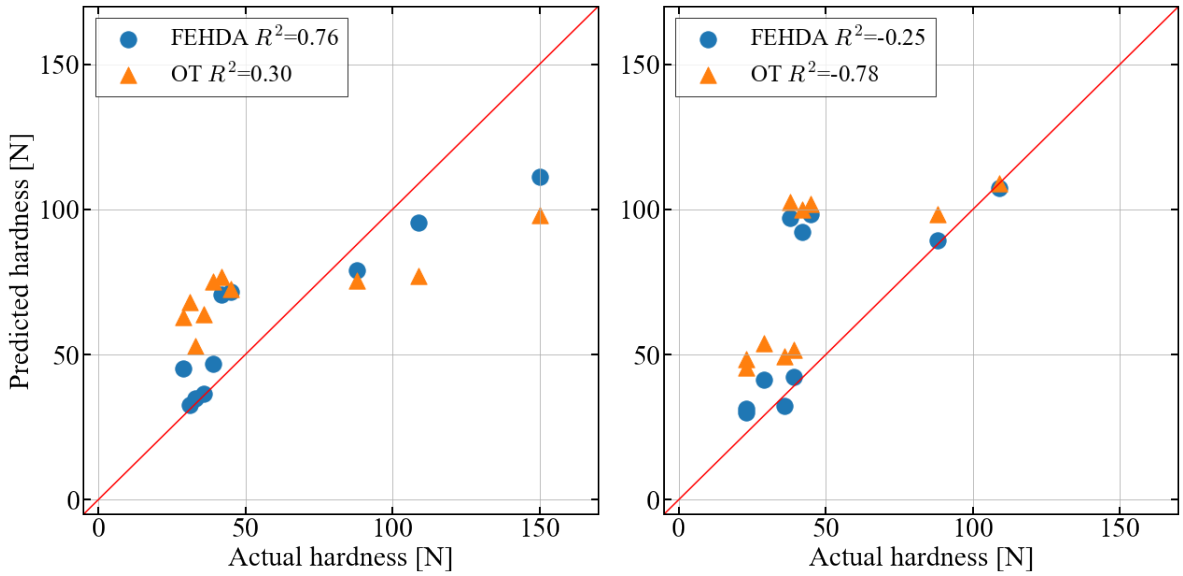

Fig. 4. Actual and predicted hardness of $D_{\text{test}}^{(\text{t})}$ by RFR models using FEHDA and OT for Splits 7 (left) and 9 (right).

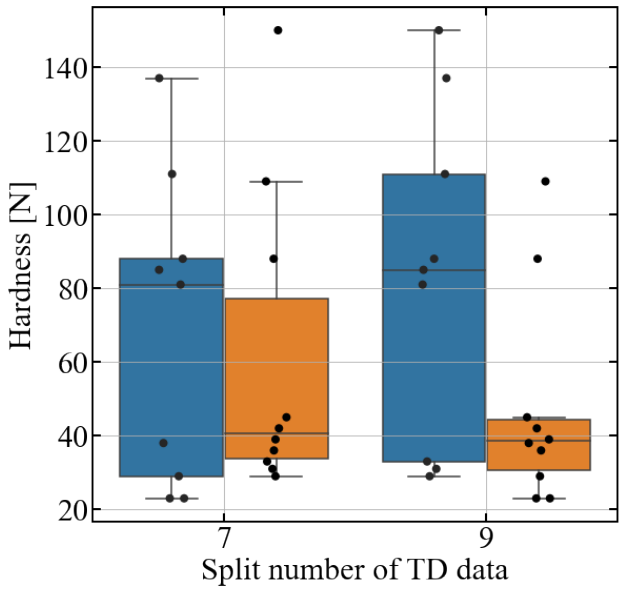

Fig. 5. Tablet hardness of the samples in $D_{\text{train}}^{(\text{t})}$ (blue) and $D_{\text{test}}^{(\text{t})}$ (orange) in Splits 7 and 9.

samples below 40 N, which may have contributed to the improved prediction performance. While both the SD data and $D_{\text{train}}^{(\text{t})}$ cover all samples in $D_{\text{test}}^{(\text{t})}$ in Split 9, one sample remains uncovered in Split 7. Even though the prediction performance improved in the extrapolation region (Ω_{ext}) in Split 7, it is still worse than that in the interpolation region (Ω_{int}), with RMSEs of 38.8 N in Ω_{ext} and 15.3 N in Ω_{int}, respectively.

Extrapolation causes the degradation of prediction performance, and other factors also play a role. In Split 9, extrapolation no longer exists; however, the prediction performance is worse than in Split 7. This is mainly due to the samples whose hardness is between 35 and 50 N.

For these samples, FEHDA does not significantly improve the prediction performance, although both the SD data and $D_{\text{train}}^{(\text{t})}$ contain samples with the hardness similar to those in $D_{\text{test}}^{(\text{t})}$, as shown in Figures 1 and 5. The values of R^2 in training were 0.964 and 0.980 in Splits 7 and 9, respectively. These results eliminate the possibility that the training did not work successfully. The potential reason for the poor prediction performance of these samples is explained below. Figure 6 shows the relationship between the main-compression force and hardness in the CDC and BDC processes. A common characteristic of the samples with poor prediction performance is that both the API content and the main-compression force are high. The total number of samples with the API content of 15% and the main-compression force of 20 N from the CDC process is three. One of them is included in $D_{\text{train}}^{(\text{t})}$ in Split 7, and none of them are included in $D_{\text{train}}^{(\text{t})}$ in Split 9. As shown in Figure 6, when the API content is 10% or 15%, as the main-compression force increases, the hardness initially increases and then decreases in both processes. As shown in Figure 4, the predicted values for these samples are higher than the actual values. This suggests that, although a similar trend is observed in the SD data, this knowledge has not been transferred from SD to TD due to the scarcity of such samples in $D_{\text{train}}^{(\text{t})}$.

A possible issue in transfer learning is that if SD and TD are not similar, transfer learning can fail and even lead to negative transfer (Rosenstein et al., 2005), where using SD data deteriorates prediction performance. In the present results, transfer learning for hardness from BDC to CDC did not cause negative transfer; however, it is necessary to verify whether it works for other CQAs. Moreover, as discussed above, even if both domains are similar, knowledge will not be transferred unless the data properly reflects their similarity. Therefore, it is essential to establish an

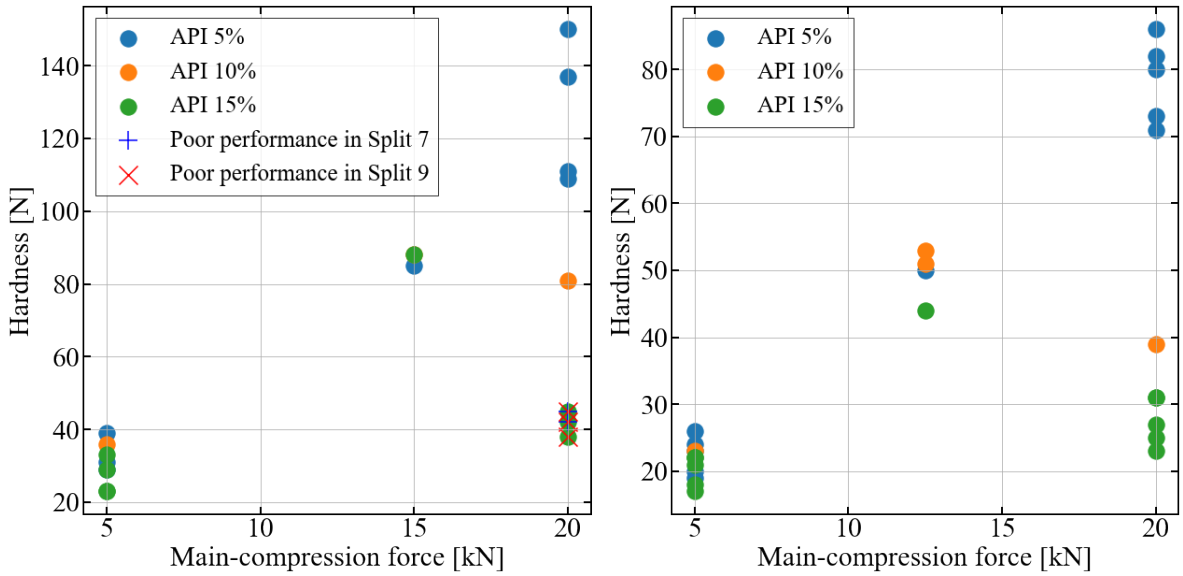

Fig. 6. The relationship between the main-compresssion force and hardness in CDC (left) and BDC (right) processes.

experimental design that supports appropriate data acquisition in TD.

4. CONCLUSIONS

This study has highlighted the potential of applying heterogeneous transfer learning (TL) to facilitate the transition from batch to continuous manufacturing in the pharmaceutical industry. The frustratingly easy heterogeneous domain adaptation (FEHDA) approach was applied to leverage data from the batch direct compression (BDC) process in predicting critical quality attributes (CQAs) for the continuous direct compression (CDC) process. Models were built using partial least squares regression (PLSR), Gaussian process regression (GPR), and random forest regression (RFR). The models outperformed models trained solely on CDC data, even in the regions with limited or no BDC data. The best result was achieved with the RFR model, which showed a reduction in RMSE by 9.36 N. Nevertheless, we also identified regions where prediction accuracy plateaued despite increased sample size. The potential reason is that because the samples in these regions do not have enough similar samples in training data from the target domain, the knowledge from the source domain was not transferred effectively. Further investigation is required to understand this underlying limitation. Moreover, our future research will focus on applying the models derived from this approach to optimize experimental design.

ACKNOWLEDGEMENTS

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123 and by JSPS KAKENHI Grant Number JP21H01704.

REFERENCES

Bang, S.H., Ak, R., Narayanan, A., Lee, Y.T., and Cho, H. (2019). A survey on knowledge transfer for manufacturing data analytics. *Comput. Ind.*, 104, 116–130.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1), 5–32.

Food and Drug Administration (2019). Quality considerations for continuous manufacturing guidance for industry.

Geladi, P. and Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, 185, 1–17.

Jones, B. and Nachtsheim, C.J. (2011). A class of three-level designs for definitive screening in the presence of second-order effects. *J. Qual. Technol.*, 43(1), 1–15.

Kobayashi, S., Miyakawa, M., Takemasa, S., Takahashi, N., Watanabe, Y., Satoh, T., and Kano, M. (2022). Transfer learning for quality prediction in a chemical toner manufacturing process. *Comput. Aided Chem. Eng.*, 49, 1663–1668.

Pan, S.J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10), 1345–1359.

Rasmussen, C.E. and Williams, C.K.I. (2005). *Gaussian processes for machine learning*. The MIT Press.

Rosenstein, M.T., Marx, Z., Kaelbling, L.P., and Dietterich, T.G. (2005). To transfer or not to transfer. In *NIPS*.

Yaginuma, K., Matsunami, K., Descamps, L., Ryckaert, A., and De Beer, T. (2024). Hybrid modeling of T-shaped partial least squares regression and transfer learning for formulation and manufacturing process development of new drug products. *Int. J. Pharm.*, 662, 124463.