2022년 2월

석사학위논문

인공지능 기반 시추공 물리검층 자료의 예측 모델 연구

조선대학교 대학원

첨단에너지자원공학과

이 득 선

인공지능 기반 시추공 물리검층 자료의 예측 모델 연구

A Study on the prediction model for well-log data based on the Artificial Intelligence

2022년 02월 25일

조선대학교 대학원

첨단에너지자원공학과

이 득 선

인공지능 기반 시추공 물리검층 자료의 예측 모델 연구

지도교수 장 일 식

이 논문을 공학 석사학위신청 논문으로 제출함

2021년 10월

조선대학교 대학원

첨단에너지자원공학과

이 득 선

이득선의 석사학위논문을 인준함

위원장 조선대학교 교 수 강 성 승 (인)

위 원 조선대학교 부교수 장 일 식 (인)

위 원 조선대학교 부교수 최 태 진 (인)

2021년 12월

조선대학교 대학원

목 차

List of Tables

List of Figures

ABSTRACT

A Study on the prediction model for well-log data based on the Artificial Intelligence

Lee, Duekseon
Advisor : Prof. Jang, Il Sic, Ph.D
Department of Advanced Energy & Resources Engineering
Graduate School of Chosun University.

Well log analysis data plays an important role in determining the reserves estimation of oil and gas, geophysical characteristics of reservoir, and wellbore stability. Sonic log is used to obtain the porosity along with the density and neutron log, and is also used to confirm the discontinuous rock formations or structures of the strata. However, there were frequent cases in which it was difficult to fully acquire the sonic log due to economic problems such as cost, defects in equipment, loss in the recording or transmission/reception stage, and wellbore problems. To solve this problem, studies were attempted to predict the sound wave detection using empirical correlation. However, the method using the empirical formula has limitations in its application in that the accuracy varies depending on a specific rock formation and geographic area.

Recently, various attempts have been made using machine learning and deep learning techniques. Among machine learning, random forest is an ensemble method that learns from multiple decision trees as an advanced technique of the decision trees. Decision trees have the advantage of having high explanatory power for the data, while have some problems of not having high

predictive ability and poor accuracy due to overfitting. Random forest is a model that compensates for these shortcomings. Through bagging, data of the same size as the original data are arbitrarily extracted and generated, and decision trees are constructed based on this. It is a method to make a final prediction by combining observations through multiple decision tree models composed of randomly selected variables with a majority vote or average value. Random forests are being used for various problems such as classification and regression.

In this study, a predictive study was performed on the unmeasured sonic log of adjacent boreholes using well log data from Volve oil field provided to the public, The random forest model, which shows good performance through an ensemble of decision trees, was applied to model construction. The input variables to be trained in the model were selected through correlation analysis, and to evaluate the reliability of the well log prediction, SVDD (support vector data description) was used to classify and visualize the areas in terms of the prediction reliability. As a result, the methodology presented in this study was confirmed that it was highly useful as a method to identify high-reliability zones in well log prediction.

제1장 서론

현대사회에서 4차 산업혁명을 주도하는 인공지능 기술은 다양한 산업에서 적용되고 끊임없이 연구되고 있다. 인공지능 기술의 하위집합인 머신러닝(machine learning)은 데이터를 기반으로하여 학습을 통해 정확도를 점차 향상시키는 방식으로 인간의 학습 능력과 같은 기능을 모사한 컴퓨터 기술이다. 이러한 머신러닝 기법은 석유가스개발 분야에서도 다양하게 이용되고 있으며 EUR 예측(오현택, 2019), 시뮬레이션 입력자료 예측(Zerrouki 등, 2014), 물리검층(well log) 데이터 복원(Alizadeh 등, 2012) 등에 적용하기 위한 연구도 계속 수행되고 있다.

시추공의 물리검층 데이터는 석유가스의 원시부존량 및 해당 지역의 유정에 대한 안정성, 지구물리학적 특성을 파악하는데 있어서 중요한 역할을 하고 있다. 그 중 음파 검층(sonic log)은 밀도 검층(density log) 및 중성자 검층(neutron log)과 더불어 공극률을 구하는데 주로 이용되며(Raymer 등, 1980), 지층의 불연속적인 암상이나 구조를 확인하기 위해 사용되기도 한다(Miller 등, 1990). 또한, 음파 검층은 셰일층에서의 급격한 음파 주시(sonic transit time) 증가특징을 활용하여 압력이 과도하게 발생하는 지점(overpressure zones)을 파악할 수 있다(Walls 등, 2000). 그러나 비용과 같은 경제적인 문제나 장비의 결함, 기록이나 송수신 단계에서의 손실, 공저에서 발생하는 문제 등과 같은 이유로 음파 검층을 온전히 획득하기 어려운 경우가 빈번하게 발생하였다(Onalo 등, 2018). 이러한 문제를 해결하기 위해 Bailey 등 (2012)과 Hossain 등 (2012)은 경험식(empirical correlation)을 이용하여 음파 검층을 예측하는 연구를 시도하였다. 그러나 경험식을 이용한 방법은 특정한 암상 및 지리적 영역에 따라 정확도가 달라진다는 점에서 적용의 한계가 발생하였다.

최근에는 머신러닝과 딥러닝(deep learning) 기법을 이용하여 다양한 시도가 이루어져왔다. Miah 등 (2020)은 shear sonic velocity를 예측하는 모델을 구축하기 위해 LS-SVM(least-squares support vector machine)을 이용하였다. 입력자료로는 RHOB(density log), DT(sonic log), NPHI(neutron log)를 이용하였으며 각 변수들이 예측하고자 하는 값에 대해 영향을 미치는 정도에 대하여 분석하였다. 이 연구에서 개발된 모델은 북해와 Niger Delta basins의 두 필드 데이터의 자료를 사용하여 예측에 대한 정확도와 신뢰도에 대해서 검증하였다.

Zeng 등 (2021)은 베이즈 이론과 grated recurrent unit에 기반한 불확실성 분석과 비선형 물리검층 데이터 예측 모델을 개발하였다. 제안된 방법은 모델과 데이터의 불확실성을 동시에 분석할 수 있다는 장점이 있다. 예측 모델은 입력자료로 GR(gamma ray), SP(spontaneous potential), RHOB, RD(deep resistivity)를 입력받아 DT를 예측하였으며 기존의 딥러닝 기반의 recurrent neural network 모델의 결과보다 더 좋은 성능을 보였다.

Otchere 등 (2021)은 Volve 유전의 유체투과도, 공극률, 물 포화도를 예측하기 위해 gradient boosting regressor를 이용하여 회귀 분석 모델을 구축하였다. 모델의 예측 성능을 향상시키기 위하여 8가지의 특성 선택(feature selection) 기법을 적용한 후 결과를 비교하였으며, 그 중 랜덤 포레스트(random forest), SelectKBest, Lasso 규제로 변수를 선택한 모델이 가장 좋은 성능을 보여주었다. 해당 연구를 통해 자료에서 주어진 모든 변수를 이용하는 방식보다 예측하고자 하는 인자를 잘 설명할 수 있는 변수만을 이용하였을 때 더 정확하고 효율적인 예측이 가능하다는 것을 확인하였다.

Jian 등 (2020)은 중국에 위치한 Daqing 유전의 데이터를 이용하여 미측정된 밀도 검층을 예측하는 모델을 구축하였다. 해당 연구에서는 ANN(artificial neural network), SVM(support vector machine) 등 기존에 자주 사용되었던 기법들은 비선형 관계를 가지는 데이터들을 정확히 반영하기 어렵다는 것을 지적하며 이를 해결하고자 deep neural network, extremely gradient boosting, 랜덤 포레스트, 다중선형 회귀, extra trees regressor 등의 모델을 적용할 것을 제안하였다. 또한 제안한 모델들의 결과를 모두 결합하여 최종 결과를 도출하는 방법을 이용하면 가장 좋은 예측 성능을 얻을 수 있다는 결론을 제시하였다. 그러나 해당 연구에서 제안한 결과를 모두 종합하는 방법은 오차를 줄이는 효과 대비 수행 시간이 길어 적용 효율성이 낮다는 한계가 있다.

Feng 등 (2021)은 머신러닝 기법 중 분류 및 회귀 분석에서 좋은 성능을 보이는 랜덤 포레스트를 이용하여 DTS(shear sonic log)를 예측하는 모델을 생성하였다. 예측값의 불확실성을 정량화하기 위하여 신뢰구간을 나누어 분석하였으며, 이미 알려진 데이터의 예측을 수행하여 실제값과 비교하였고, 이를 통해 모델의 정확성을 검증하였다. 랜덤 포레스트 모델의 파라미터를 최적화하기 위하여 그리드 탐색(grid search) 방식을 적용하였으며, 다수의 입력 변수가 존재할 경우 다중공선성(multicolinearity)의 문제를 해결할 수 있는 PCA(principal component analysis)를

통해 입력변수의 차원을 축소하여 생성한 모델의 결과를 비교하였다. PCA를 통해 축소한 입력 변수를 이용한 결과와 전체 입력 변수를 이용한 예측 결과가 크게 다르지 않았다. 이 연구에서는 실제로 미측정된 지역의 데이터를 예측한 것이 아니라 이미 알고 있는 데이터의 구간을 임의로 미측정되었다고 가정하여 수행하였으며 예측 구간을 선정하는 기준이 명확히 제시되지 않았다.

위 선행연구들은 동일한 시추공 물리검층 데이터만을 이용하여 학습 및 예측을 진행하였으므로 다른 시추공 데이터를 예측하는 것보다 비교적 예측이 쉬울 것으로 예상되며, 다른 시추공의 미측정된 물리검층 예측에 대한 검증이 어려운 한계가 있다.

이 연구에서는 공공에 제공된 Volve 유전 데이터의 물리검층 자료를 학습한 후 인접 시추공의 미측정된 음파 검층에 대한 예측 연구를 수행하였다. 모델 구축에는 의사결정나무(decision tree)의 앙상블(ensemble)을 통해 좋은 성능을 보여주는 랜덤 포레스트 기법을 적용하였다. 모델에 학습시킬 입력 변수는 상관관계 분석을 통하여 선정하였으며, 물리검층 예측에 대한 신뢰도를 평가하기 위해 SVDD(support vector data description)를 이용하여 예측 신뢰도에 따른 영역을 구분하여 시각화 시켰다.

제2장 이론적배경

제1절 머신러닝(Machine Learning)

머신러닝이란 인공지능의 한 분야로서 데이터를 기반으로 패턴을 학습하고 결과를 예측하여 의사결정을 하는 것을 말한다.(김은미, 2020) 머신러닝은 1950년대 인공지능이라는 개념으로 시작하여 연구가 진행되어오다 1970년대부터 1980년대까지 10년동안 과적합(Overfitting), XOR 분류 문제 등에 부딪혀 오랜 기간 침체기를 겪었다. 이후 인공 신경망과 더불어 다층 퍼셉트론(multilayer perceptron), 의사결정 나무, SVM 등 다양한 알고리즘의 개발이 이루어졌다.

머신러닝은 학습 방법에 따라 크게 지도학습(supervised learning), 비지도 학습(unsupervised learning), 강화 학습(re-enforced learning)으로 분류할 수 있다. 지도학습은 학습 데이터(training data)가 입력 데이터(input data)와 출력 데이터(output data)의 쌍으로 구성되어야 하는 기법으로, 학습은 입력에 대한 머신러닝 모델의 출력과 학습 데이터의 출력의 차이가 줄도록 모델을 수정해 가는 과정을 나타낸다(김성필, 2016). 지도학습은 분류(classification)와 회귀(regression) 문제로 나눌 수 있다. 분류 모델은 학습 데이터의 출력 데이터로 범주를 학습시켜 입력 데이터가 어떠한 범주에 속하는지 찾아내는 문제이며, 회귀 모델은 입력 데이터와 예측하고자 하는 값을 출력 데이터로 모델을 학습시켜 새로운 자료를 입력 받았을 때 학습한 알고리즘을 바탕으로 어떤 값을 예측하는 모델이다.

비지도 학습은 기계학습 분야에서 결과 값이 알려진 상황에서의 예측모델인 지도학습과는 다르게 결과 값이 정확히 주어지지 않은 자료에 대한 분석을 말한다. 이러한 학습 과정을 통해 데이터들의 특징 및 성격을 파악할 수 있으며, 비지도 학습에서의 데이터 마이닝은 방대한 양의 데이터로부터 알고리즘을 통해 유용한 패턴을 찾을 수 있도록 해준다(Witten et al, 2011). 비지도 학습을 통한 데이터마이닝 기법으로는 k-평균군집, 계층적 군집, 혼합 분포군집을 포함하는 다양한 군집분석과 주성분 분석, 요인분석, 독립성분 분석 등이 포함된다(임성태, 2019).

1. 앙상블(Ensemble)

앙상블은 일련의 예측기(즉, 분류나 회귀 모델)로부터 예측 정보를 수집하여 종합함으로써 가장 좋은 모델 하나보다 더 좋은 예측 결과를 얻는 방법이다. 앙상블의 예를 들면 훈련 세트로부터 무작위로 각기 다른 서브셋을 생성하여 의사결정 트리를 훈련시킬 수 있다. 또한 그 성능이 뛰어나 머신러닝 경연대회에서 주로 사용되었고, 특히 넷플릭스 대회(netflixprize)에서 가장 인기 있는 방법이었다. (박해선, 2021)

앙상블의 대표적인 유형으로는 배깅(bagging), 부스팅(boosting)이 있다. 배깅은 boostrap aggregating의 약자로 앙상블을 이용한 머신러닝 방법이며 Breiman(1994)에 의해서 제안되었다. 부트스트랩은 데이터 내에서 중복을 허용하여 반복적으로 샘플을 추출하는 방법 중 하나이다. 또한 여러 개의 독립적인 모델을 생성하는데 이 과정에서 각각의 모델들은 서로에게 영향을 주지 않는다. 배깅은 이렇게 생성된 여러 모델의 결과를 이용하여 전체 모델의 평균(average)을 계산하거나 과반수 투표(voting)를 실시하여 최종 결과를 도출해내는 방법이다.

부스팅은 약한 학습기(weak learner)를 여러개 연결하여 강한 학습기(strong learner)를 만드는 앙상블 방법을 말한다. 이전 학습에 대해 잘못 예측된 데이터에 가중치(weight)를 부여하여 오차를 보완해 나가는 방법이다. 부스팅은 주로 에이다부스트(AdaBoost)와 그래디언트 부스팅(gradient boosting)으로 구분된다. 에이다부스트는 이전 모델이 잘못 예측한 데이터에 가중치를 더 높여, 새로운 모델은 업데이트된 가중치를 사용하여 점차 더 정확한 예측을 하는 방식이다. 그래디언트 부스팅은 에이다부스트와 마찬가지로 앙상블에 이전까지의 오차를 보정하도록 예측기를 순차적으로 추가한다. 하지만 샘플의 가중치를 수정하는 에이다부스트와는 달리 이전 예측기가 만든 잔여 오차(residual error)를 새로운 모델에 학습시키는 방식이다.

2. 랜덤 포레스트(random forest)

랜덤 포레스트는 다수의 결정 트리들을 학습하는 앙상블 방법으로 의사결정나무의 심화된 기법이다. 의사결정나무는 데이터에 대한 설명력이 높다는 장점이 있으며 예측 능력이 높지 않은 문제점과 과적합으로 인하여 정확도가 떨어지는 단점이 있다. 랜덤 포레스트는 이와 같은 단점을 보완한 모형으로 배깅을 통해 임의로 원본 데이터와 동일한 크기의 자료들을 추출 및 생성하여 이를 기반으로 의사결정나무들을 각각 구성한다. 이후 무작위로 선택된 변수들로 구성된 여러 의사결정나무 모형을 통한 관측치를 다수결이나 평균값으로 결합하여 최종적으로 예측하는 방식이다(김준석, 2021). 랜덤 포레스트는 분류, 회귀 등 다양한 문제에 활용되고 있다.

랜덤 포레스트는 CART(classification and regression tree) 알고리즘을 기반으로 구성되어 입력 데이터와 출력 데이터의 종류에 제약이 거의 없으며 알고리즘 특성상 잡음이나 이상치에 받는 영향이 적다. 랜덤 포레스트는 다양한 분야에 적용되어 왔으며, 많은 알고리즘 중 성능이 뛰어난 모델 중 하나로 평가되고 있다(이예지, 2019). Figure 2.1은 전체 데이터셋에서 다수의 결정 트리들로 최종 결과를 도출하는 과정을 설명한 것이다.

랜덤 포레스트는 생성된 모델의 편향(bias)를 유지하면서 분산(variance)를 감소시킨다는 장점이 있으며, 단점으로는 샘플 추출 시 중복 추출을 허용하므로 특정 샘플은 추출이 안되는 편향성의 문제가 발생할 수 있다. 이를 보완하기 위해 oob(out-of bag) 샘플을 이용하여 학습 시에 추출되지 않는 데이터들의 예측 결과를 확인하여 검증이 가능하다. 여기서 oob란 전체 데이터 중 추출되지 않는 샘플로서 전체 데이터 중 약 37%가 이에 해당한다 (Geron, 2019).

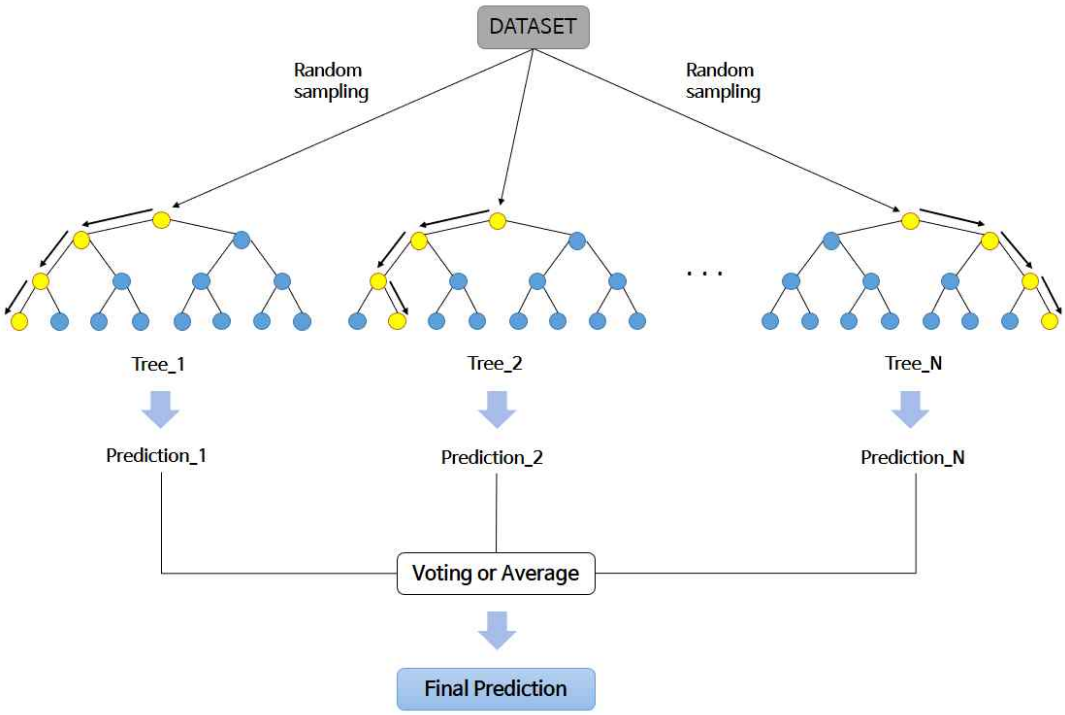


Figure 2.1 Example scheme of Random Forest.

3. One-Class SVM

One-class SVM(one-class support vector machine)은 주어진 데이터를 잘 설명할 수 있는 최적의 서포트 벡터를 구하고 이 영역 밖의 데이터들을 이상치로 간주하는 방식의 기법이다(Scholkopf 등, 1999). 데이터들을 좌표 축으로 옮긴 후, 원점과의 거리를 기준으로 선을 그어 이상치와 정상 데이터를 구분한다. 정상 데이터는 최대한 원점에서 멀어지도록 맵핑한다. 일반적으로 이진분류(binary-classification) 문제에 자주 사용되는 SVM은 초평면(hyper plane)을 기준으로 데이터를 나누어 2개의 범주로 분류하지만, One-class SVM은 조건에 따라서 1개의 범주와 이상치로 분류한다. One-Class SVM의 목적함수는 아래의 식 (1)과 같다.

$$\min_{w,\xi_i,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \qquad (1)$$

$$\text{subject to } w \bullet \Phi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0 \text{ for all } i = 1,2,\cdots,n$$

$\|w\|^2$: regularizer of the hyperplane w

ρ : distance from origin to hyperplane

ξ_i : penalty term for soft magin

ν : hyper parameter for controlling the trade-off

n : number of data

Φ : function to mapping original data into feature space

4. Support Vector Data Description

데이터 마이닝 기법 중 하나인 SVDD는 이상치 벡터를 탐지하기 위해 Tax 등(2004)이 처음으로 제시한 단일 클래스 분류기법이다. One class SVM의 목적이 원점에서 가장 멀리 떨어진 초평면 경계(hyperplane boundary)를 찾는 것이라면 SVDD의 목적함수는 정상 데이터를 감싸 안는 최소 최적의 구(hypershpere)를 찾는 boundary를 탐색하는 것이다. (안성호, 2018) 이를 통해 정상 데이터와 비정상 데이터를 나누는 것이 가능하다. SVDD의 목적 함수는 아래의 식과 같다.

$$\min_{R,a,\xi} R^2 + C\sum_{i=1}^{n}\xi_i \tag{2}$$

$$\text{subject to } \| \Phi_k(x_i) - a \|_{F_k}^2 \leq R^2 + \xi_i,\ \xi_i \geq 0 \quad \text{for all } i = 1,2,\cdots,n$$

R : radius

a : center

ξ_i : penalty for soft majin

ν : hyper parameter for controlling the trade-off

C : hyper parameter for controlling the error

Φ_k : function to mapping original data into feature space

F_k : feature space

구를 생성하는 방식으로는 커널을 이용하는데, 커널 함수로는 다항식 커널(polynomial kernel)과 가우시안 커널(gaussian kernel)이 있다. 그 중 주로 사용되는 가우시안 커널(gaussian kernel)의 식 (3)은 아래와 같다.

$$K(x_i,x_j) = \exp\left(\frac{-\| x_i - x_j \|^2}{s^2}\right) \tag{3}$$

x_i, x_j : data points

s : kernel width

Figure 2.2는 s에 따른 경계면의 형성효과를 설명한 것이다. s는 커널의 너비(kernel width)를 나타내며 값이 커질수록 데이터를 구 모양으로 경계면을 만들고, 값이 작을수록 데이터에 최대한 가깝게 경계면을 만든다.

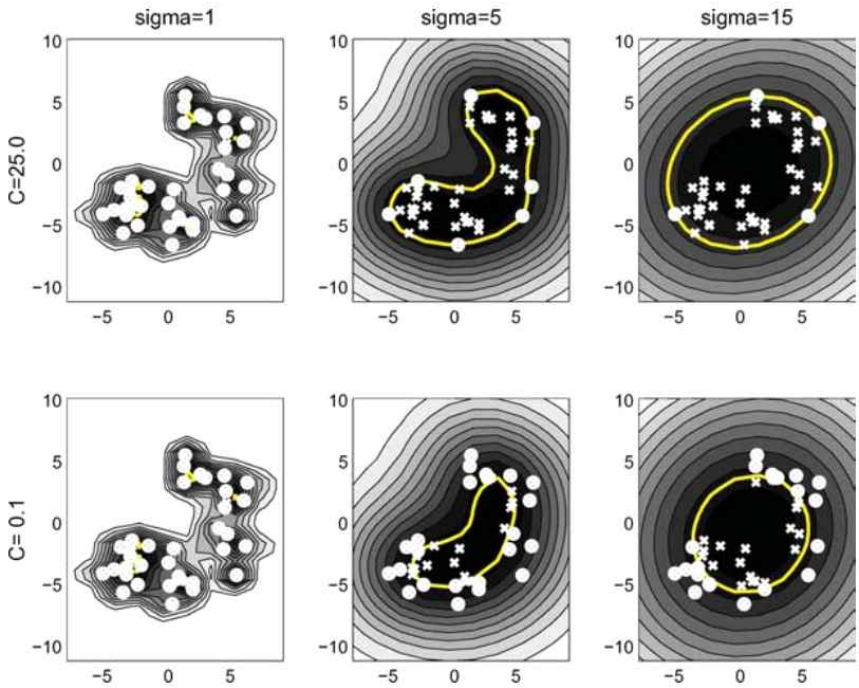


Figure 2.2 Data description trained on a banana shaped data set(Tax et al., 2004).

제3장 음파 검층 로그 예측을 위한 머신러닝 모델 구축 연구

제1절 대상 데이터

본 연구에서 사용한 음파검층 데이터는 노르웨이 Volve 유전의 시추공 자료이다. 15/9-F-1A, 15/9-F-1B, 15/9-F-1C, 15/9-F-11A, 15/9-F-11B의 물리검층 데이터를 사용하였으며, 데이터는 LAS Log ASCII 표준 파일 형식이다. 물리검층 데이터의 측정 시작위치와 종료위치는 시추공마다 다르며 Depth를 기준으로 0.1 m 간격으로 측정되었다. 해당 데이터는 경도 및 위도에 대한 정보를 포함하고 있으며, 'Statoil' 사에서 측정한 데이터로 현재는 'Equinor' 사로 명칭이 변경되었다. 자료의 다수의 구간에서 '-999.25' 로 기록되어 있는 수치들을 확인할 수 있는데 해당 값은 null 값으로 알 수 없거나 측정하지 못한 값을 나타낸다. 아래 Figure 3.1에서 Volve 유전의 위치를 확인할 수 있다.

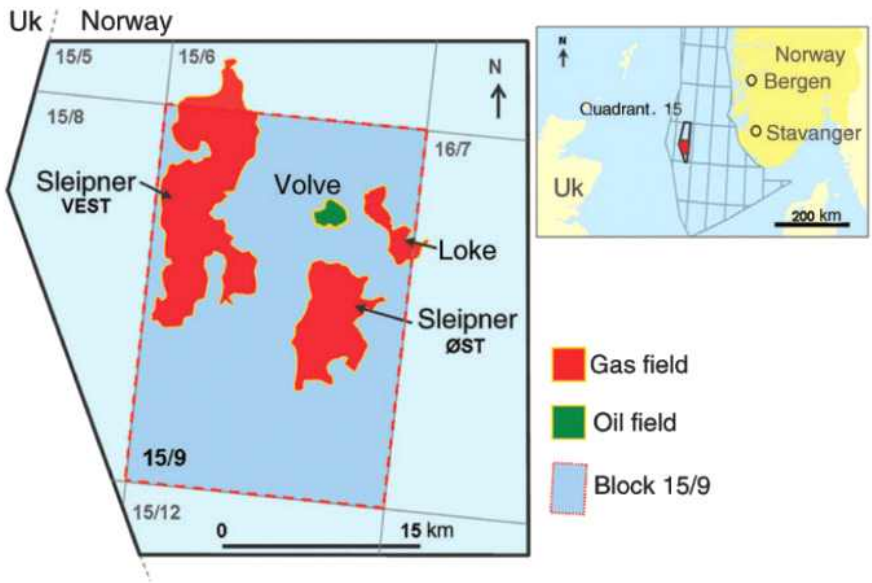


Figure 3.1 Location of Volve field in the North Sea (Ravasi et al. 2015).

각 데이터는 공통으로 측정된 검층 항목과 이외의 항목을 포함하고 있다. 공통으로 포함된 특성은 다음의 Table 3.1과 같다.

Table 3.1 Common log type list

Well name	logging type	unit	commnet
15/9-F-1A	ABDCQF01	g/cm^3	
	ABDCQF02	g/cm^3	
	ABDCQF03	g/cm^3	
	ABDCQF04	g/cm^3	
15/9-F-1B	BS	inches	Bit size
	CALI	inches	Caliper
	DRHO	g/cm^3	Density correction
15/9-F-1C	GR	API	Gamma ray
	NPHI	v/v	Neutron
15/9-F-11A	PEF	b/elec	Photoelectric factor
	RACEHM	unknown	
	RACELM	unknown	
15/9-F-11B	RHOB	g/cm^3	Bulk density
	ROP	m/hr	Rate of penetration
	RPCEHM	unknown	
	RPCELM	unknown	
	RT	ohm.m	Resistivity

추가적으로 15/9-F-1A, 15/9-F-1B, 15/9-F-11A 데이터는 음파 검층 항목인 DT, DTS가 모두 존재하며, 반면에 15/9-F-1C, 15/9-F-11B 데이터는 DT, DTS가 모두 존재하지 않았다. 따라서 음파 검층을 보유한 15/9-F-1A, 15/9-F-1B, 15/9-F-11A를 이용하여 모델을 학습시킨 후 음파 검층이 없는 15/9-F-1C, 15/9-F-11B의 음파 검층에 예측하였다.

공통으로 포함하고 있는 17개의 로그 항목 중 음파 검층을 예측하는 데 관계가 없는 3개의 검층 항목인 BS(bit size), CALI(caliper diameter), ROP(rate of penetration)는 입력자료에서 제외시켰다.

제2절 데이터 전처리(Data preprocessing)

1. 상관관계 분석을 통한 입력 변수 선정

각 입력변수의 상관성을 정량적으로 분석하기 위해 상관계수 분석을 수행하였다. 상관관계의 유형으로는 피어슨(Pearson), 켄달(Kendall), 스피어만(Spearman)이 있으며, 보편적으로 많이 이용되는 피어슨 방식을 이용하여 상관계수를 도출하였다. 피어슨 상관 계수를 구하는 식은 식 (4)와 같다. 상관관계는 데이터 간 단순 관계를 설명하는데 유용하다. 상관계수 r값이 +1 , -1에 가까울수록 상관성이 높으며, 0에 가까울수록 상관성이 낮다.

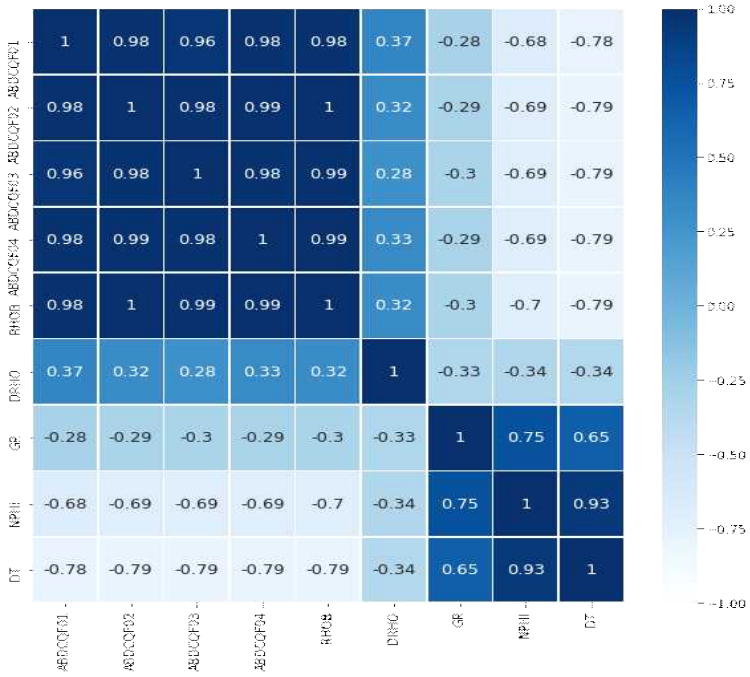
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (4)$$

n : sample size

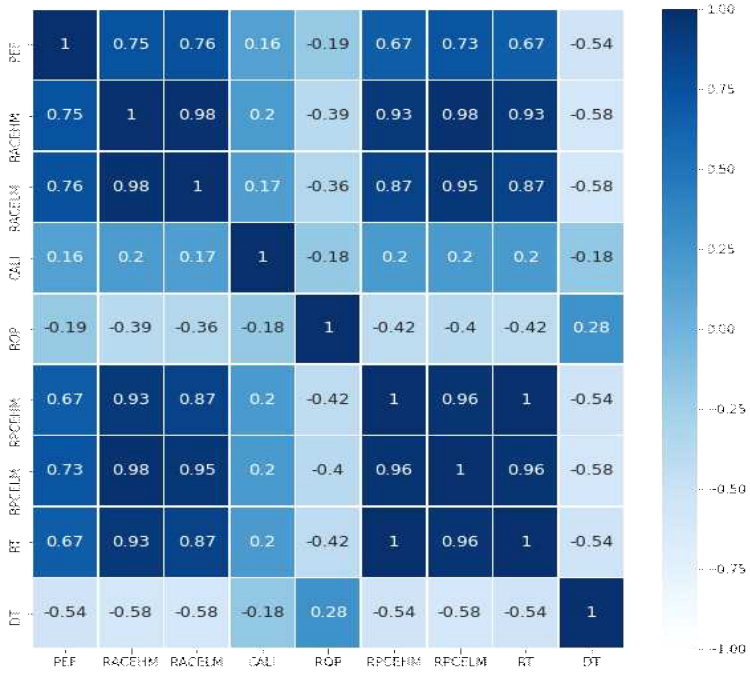
x_i : input variable sample

y_i : output variable sample

학습용 데이터인 15/9-F-1A, 15/9-F-1B, 15/9-F-11A의 검층 항목 중 14개의 변수를 대상으로 DT와의 상관관계 분석을 수행하였다. 상관관계를 분석하기 전 비저항 검층(RT)은 로그스케일로 변환하여 사용하였다. 상관관계 분석 결과는 Figure 3.2에 나타내었다.



(a)



(b)

Figure 3.2 Feature correlation coefficient using Pearson Correlation.

상관관계 분석 결과 위 Figure 3.2(a)에서 ABDCQF01, ABDCQF02, ABDCQF03, ABDCQF04, RHOB 간의 상관계수가 0.98 이상임을 확인하였다. 회귀 분석 모델을 생성할 때 예측을 위해 학습되는 입력변수 사이의 상관계수가 너무 높으면 다중공선성(multicolinearity)의 문제가 발생하여 예측 성능에 악영향을 미치게 된다. 따라서 위의 5개의 입력변수 중 RHOB만을 모델에 학습시킬 변수로 선정하였으며, 상관계수가 0.5 이상인 GR, NPHI 또한 입력변수로 선정하였다.

Figure 3.2(b)에서도 RACEHM, RACELM, RPCEHM, RPCELM, RT 간의 상관계수가 0.9 이상임이 확인되었다. 따라서 5개의 변수 중 시추공 물리 검층 시 흔히 측정되는 RT와 상관계수가 0.5 이상인 PEF를 입력 변수로 선정하였다.

따라서 변수들의 상관관계를 비교하여 최종적으로 모델에 사용할 입력 변수 및 출력 변수를 Table 3.2에 정리하였다.

Table 3.2. Input and output variable for model construction

Input variable	NPHI	Neutron Porosity
	RHOB	Bulk density
	GR	Gamma ray
	RT	Resistivity
	PEF	Photoelectric Factor
output variable	DT	Compressional sonic log

2. 이상치 제거 (outlier removing)

이상치(outlier)란 관측된 데이터의 전체적인 패턴이나 범위에서 벗어난 값을 말한다. 이상치는 모델이 의사결정을 하는데 영향을 미칠 수 있으므로 제거해 주어야 한다. 각각의 물리검층 데이터에서 이상치를 제거해주기 위해 One-class SVM 기법을 이용하였다. 이상치를 제거할 데이터는 모델 학습 시에 사용할 데이터와 예측에 사용할 미측정 로그를 포함한 데이터를 대상으로 수행하였다. One-class SVM 기법을 이용하기 위해서는 각 변수들을 식 (5)를 사용하여 정규화하였다.

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

x: each value in the data set

μ: mean

σ: standard deviation

학습에 사용할 데이터를 정규화하여 얻은 μ와 σ값을 이용하여 예측용 데이터에 그대로 적용하여 변환하였다.

Figure 3.3은 15/9-F-1A 데이터와 15/9-F-11A 데이터의 정규화 이후의 데이터 분포를 나타내는 그래프이다. 이처럼 데이터를 정규분포로 변환하여 이상치를 쉽게 파악할 수 있으며, 15/9-F-1A, 15/9-F-11A 데이터 모두 이상치가 존재하는 것을 확인하였다. 다음으로 One-class SVM를 이용하여 해당하는 이상치를 제거해주는 과정을 수행하였다.

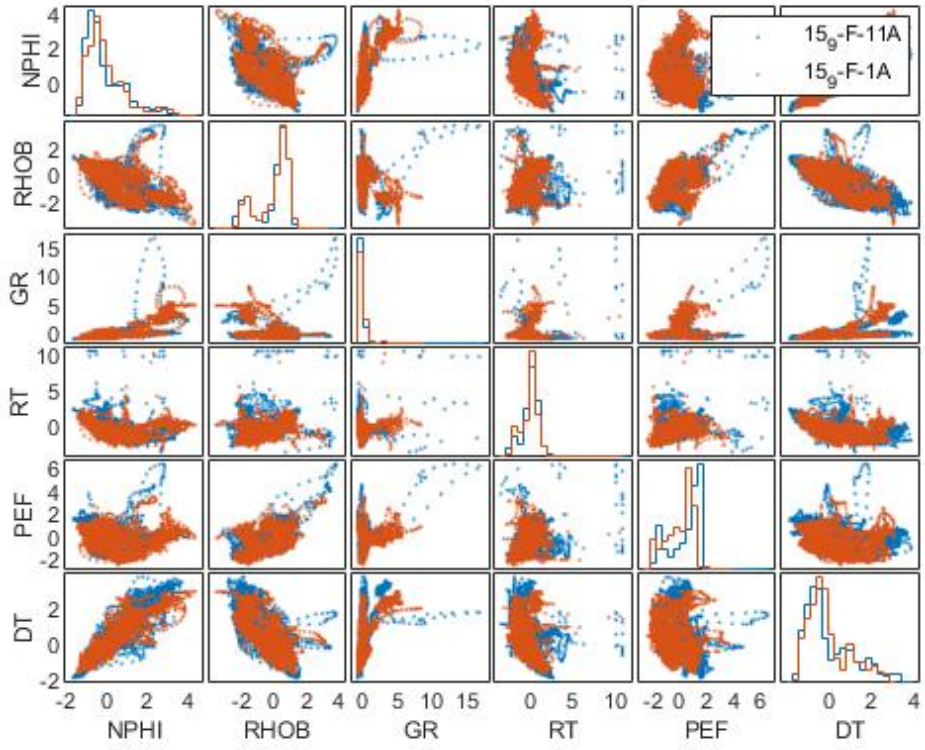


Figure 3.3 Cross plot between 15/9-F-11A and 15/9-F-1A.

학습용 데이터와 예측용 데이터를 결합하여 전체 데이터 중 이상치가 최대 10% 정도 존재한다고 가정하고 One-class SVM을 적용하여 이상치를 제거하였다. Figure 3.4는 15/9-F-1A, 15/9-F-11A 데이터의 이상치를 제거한 후 데이터 분포를 나타낸 그래프이다. 이상치 제거 전과 비교하면 각 변수들의 이상치가 적절히 제거된 것을 확인할 수 있다.

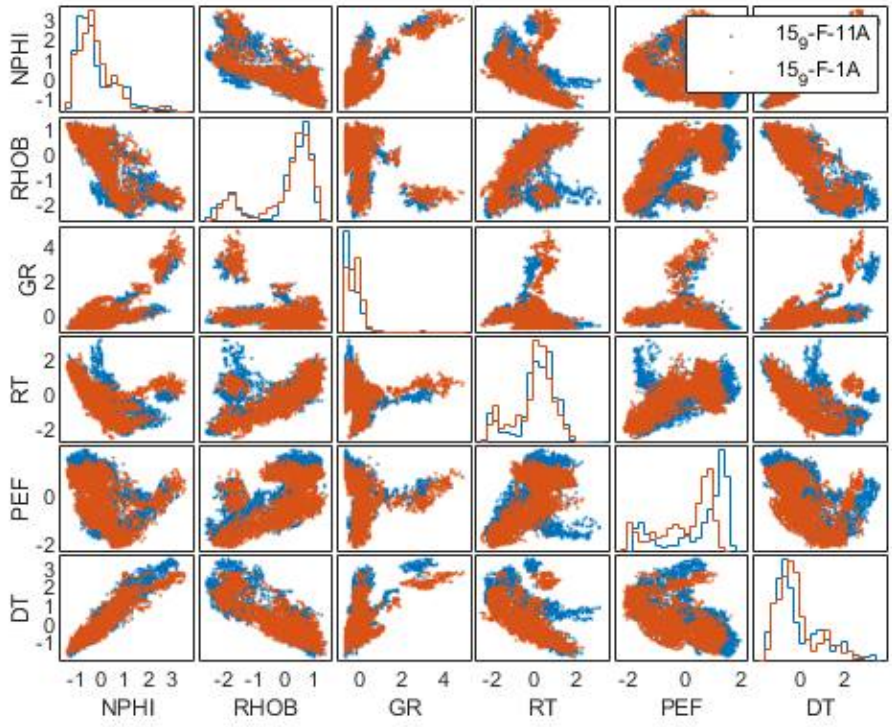


Figure 3.4 Cross plot between 15/9-F-11A and 15/9-F-1A after removing outlier.

Table 3.3은 15/9-F-1A, 15/9-F-11A의 이상치를 제거한 이후 상관계수를 계산한 결과이다. 해당 연구에서 예측 대상인 DT와 NPHI, RHOB, GR, RT, PEF 모두 상관계수가 0.6 이상으로 비교적 높은 상관관계를 나타냈으며, 그 중 NPHI와 RHOB는 각각 0.94, -0.87으로 높은 상관성을 보였다.

Table 3.3 Correlation coefficient of 15/9-F-11A and 15/9-F-1A after removing outlier

feature name	NPHI	RHOB	GR	RT	PEF	DT
NPHI	1.00	-0.80	0.76	-0.56	-0.54	0.94
RHOB	-0.80	1.00	-0.34	0.81	0.76	-0.87
GR	0.76	-0.34	1.00	-0.02	-0.20	0.63
RT	-0.56	0.81	-0.02	1.00	0.78	-0.67
PEF	-0.54	0.76	-0.20	0.78	1.00	-0.66
DT	0.94	-0.87	0.63	-0.67	-0.66	1.00

제3절 랜덤 포레스트 기법을 이용한 예측 모델 생성

본 연구에서는 미측정 시추공의 음파 검층을 예측하기 위해 학습용 데이터로 15/9-F-1A, 15/9-F-1B, 15/9-F-11A를 사용하여 랜덤 포레스트 모델을 생성하였다.

랜덤 포레스트 회귀 모델의 성능을 평가하기 위하여 RMSE(root mean squared error)를 이용하였으며 트리의 개수는 500개로 설정하였다. RMSE는 MSE(mean squared error), MAE(mean absolute error) 와 함께 회귀 모델의 성능을 평가하기 위하여 주로 사용하며 식 (6)을 통하여 계산한다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)} \tag{6}$$

y : true value

$\hat{y}$: predicted value

n : number of data

미측정 데이터의 예측에 앞서 모델의 예측 신뢰도를 평가하기 위해 15/9-F-1A의 전체 데이터를 모델에 학습시킨 후 15/9-F-1B, 15/9-F-11A의 DT 예측을 실행하였으며, 동일한 방법으로 15/9-F-1B 전체 데이터를 학습한 모델을 생성한 후 15/9-F-1A, 15/9-F-11A의 DT를 예측해 보았다. 또한, 15/9-F-11A도 동일하게 수행하였다.

다음으로는 두 개의 데이터를 합쳐 학습한 후 나머지 데이터를 예측하는 방식으로 모델을 생성하였으며, 생성한 모델을 구별하기 위해 모델 명을 'RF_학습데이터명_예측데이터명_목적_model'로 설정하였다. 예를 들어 'RF_1A_1B_training_model'은 15/9-F-1B의 DT 예측을 위해 15/9-F-1A 데이터로 학습한 모델의 학습결과를 의미하며, 'RF_1A_1B_validation_model'은 15/9-F-1A를 사용하여 학습한 이후 검증결과를, 'RF_1A_1B_prediction_model'은 15/9-F-1A를 사용하여 학습한 모델로부터 5/9-F-1B의 DT에 대한 예측결과를 의미한다. 이와 같은 방식으로 생성된 모델은 총 9개이며 각각의 예측 결과의 RMSE를 비교하였다.

1. 15/9-F-1A 예측 모델 결과

1) 15/9-F-1B를 학습자료로 사용한 경우

Figure 3.5는 15/9-F-1B 데이터를 학습하여 15/9-F-1A 데이터의 DT를 예측한 결과이다. 학습데이터에 대한 RMSE가 0.0792이며, 검증데이터의 RMSE는 0.1228로 그래프 상으로도 학습이 잘 된 것을 확인하였다. 그러나 15/9-F-1A의 DT에 대한 예측 결과의 RMSE는 0.3973이며 실제 값과 예측 값이 그래프 상으로 잘 매칭되지 않는 것으로 보아 15/9-F-1B 데이터를 단독으로 학습하여 15/9-F-1A 데이터의 DT를 예측하는 것에 한계가 있음을 알 수 있다.

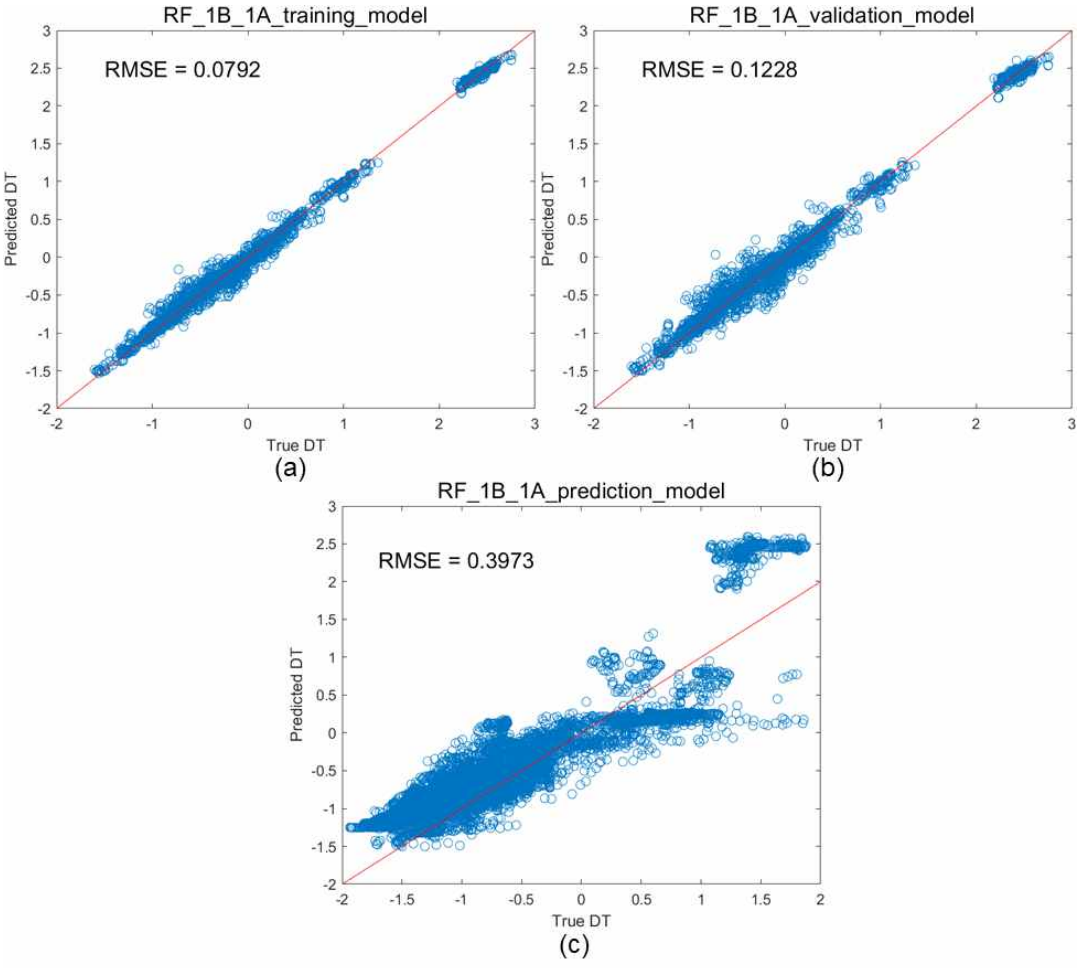


Figure 3.5 (a) Training, (b) validation using 15/9-F-1B data and (c) prediction of DT in 15/9-F-1A.

2) 15/9-F-11A를 학습자료로 사용한 경우

Figure 3.6은 15/9-F-11A 데이터를 학습한 후 15/9-F-1A의 DT를 예측한 모델의 결과이다. 모델의 학습 결과를 통해 계산한 RMSE는 0.0783이고, 검증 결과의 RMSE는 0.1191, 15/9-F-1A의 DT를 예측한 결과의 RMSE는 0.2215이다. 예측 결과 그래프를 보면 예측된 데이터의 끝부분이 실제 값보다 더 크게 예측된 것을 확인할 수 있었다. 15/9-F-1B를 학습하여 예측한 결과와 비교하면 15/9-F-11A 데이터를 학습시킨 모델이 더 좋은 성능을 보이는 것으로 나타났다.

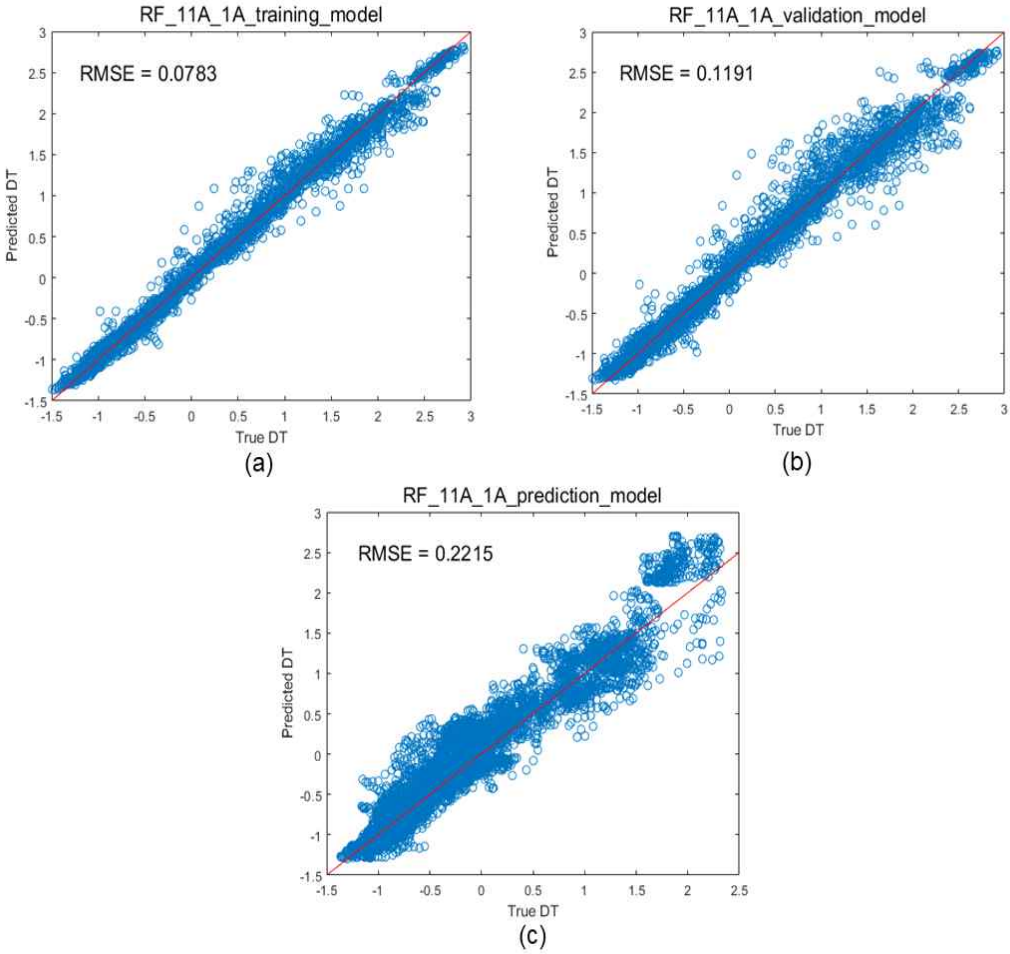


Figure 3.6 (a) Training, (b) validation using 15/9-F-11A data and (c) prediction of DT in 15/9-F-1A.

3) 15/9-F-1B와 15/9-F-11A를 학습자료로 사용한 경우

Figure 3.7은 15/9-F-1B 데이터와 15/9-F-11A 데이터를 함께 사용하여 모델을 학습한 후 5/9-F-1A의 DT를 예측한 모델의 결과이다. 모델의 학습 결과를 통해 계산한 RMSE는 0.0788이며, 검증 결과의 RMSE는 0.1200, 15/9-F-11A의 DT를 예측한 결과의 RMSE는 0.2230이다. 예측 결과의 그래프를 살펴보면 예측한 데이터의 끝부분이 실제 데이터보다 더 크게 예측되었음을 확인하였다. 해당 결과는 15/9-F-11A를 이용하여 학습시킨 모델과 성능이 비슷하며, 15/9-F-1B를 이용하여 단독으로 학습시킨 모델보다 성능이 좋았다.

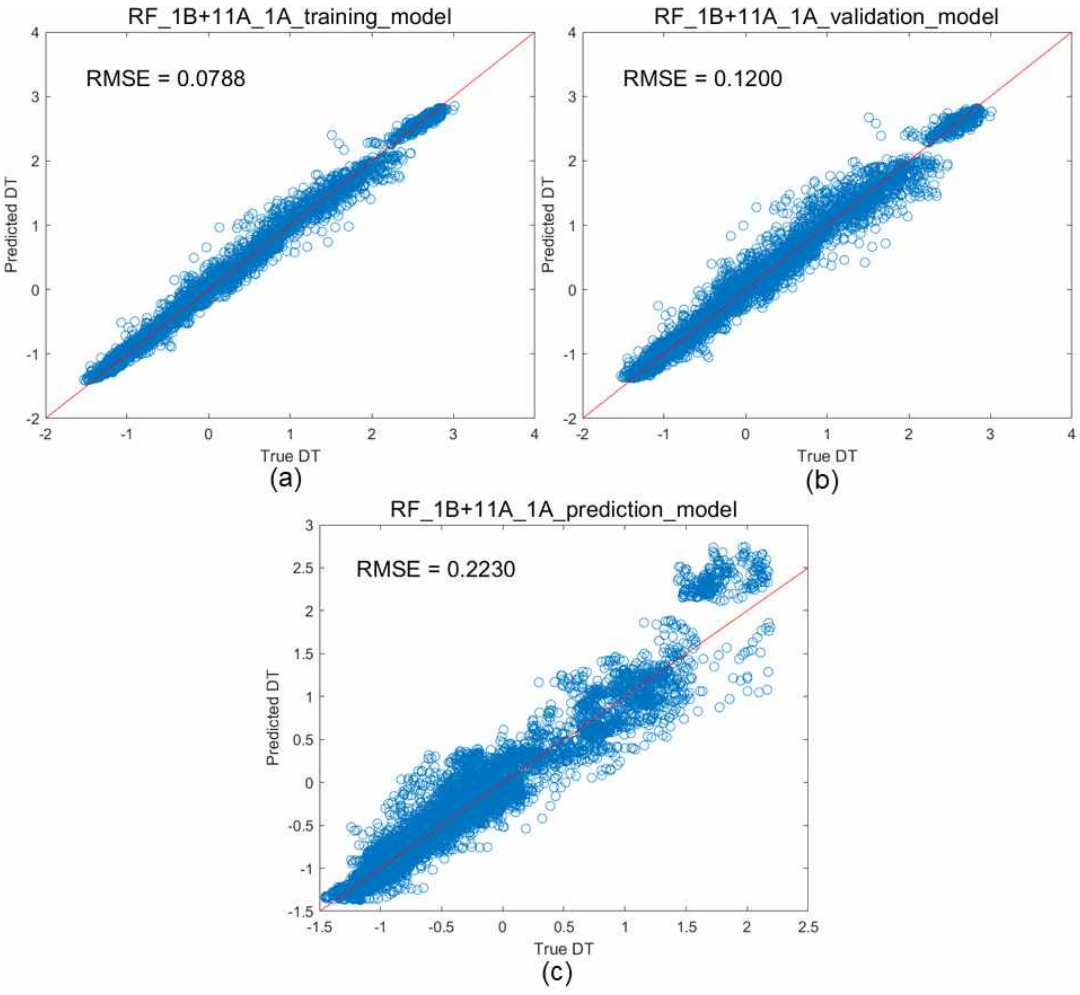


Figure 3.7 (a) Training, (b) validation using 15/9-F-1B+11A data and (c) prediction of DT in 15/9-F-1A.

2. 15/9-F-1B 예측 모델 결과

1) 15/9-F-1A를 학습자료로 사용한 경우

Figure 3.8은 15/9-F-1A 데이터를 학습한 후 15/9-F-1B의 DT를 예측한 모델의 결과이다. 모델의 학습 결과 RMSE는 0.0870이고, 검증 결과의 RMSE는 0.1342, 15/9-F-1B의 DT를 예측한 결과의 RMSE는 0.5011이다. 예측 결과 그래프를 보면 예측된 데이터의 끝부분이 실제 값보다 더 작게 예측된 것을 확인할 수 있었다. 따라서 15/9-F-1A를 단독으로 학습하여 15/9-F-1B를 예측하는 것에 한계가 있음을 알 수 있다.

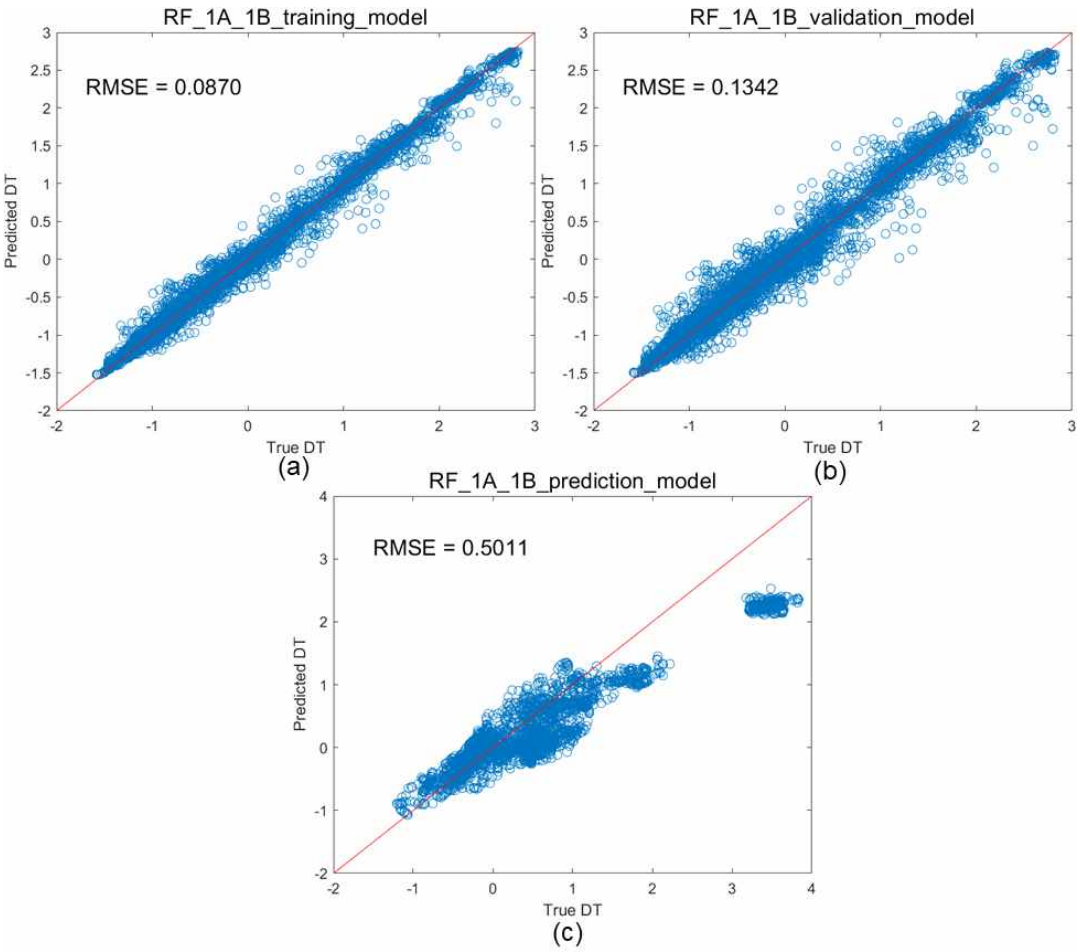


Figure 3.8 (a) Training, (b) validation using 15/9-F-1A data and (c) prediction of DT in 15/9-F-1B.

2) 15/9-F-11A를 학습자료로 사용한 경우

Figure 3.9는 15/9-F-11A 데이터를 학습한 후 15/9-F-1B의 DT를 예측한 모델의 결과이다. 모델의 학습 결과를 통해 계산한 RMSE는 0.0752이고, 검증 결과의 RMSE는 0.1146, 15/9-F-1B의 DT를 예측한 결과의 RMSE는 0.2538이다. 예측 결과 그래프를 보면 예측된 결과와 실제 결과가 비교적 잘 매칭되는 것을 확인할 수 있었다. 또한 15/9-F-1A를 단독으로 학습하여 예측한 결과의 RMSE보다 해당 예측 결과의 RMSE가 50%정도 줄어듦으로써 더 좋은 성능을 보였다.

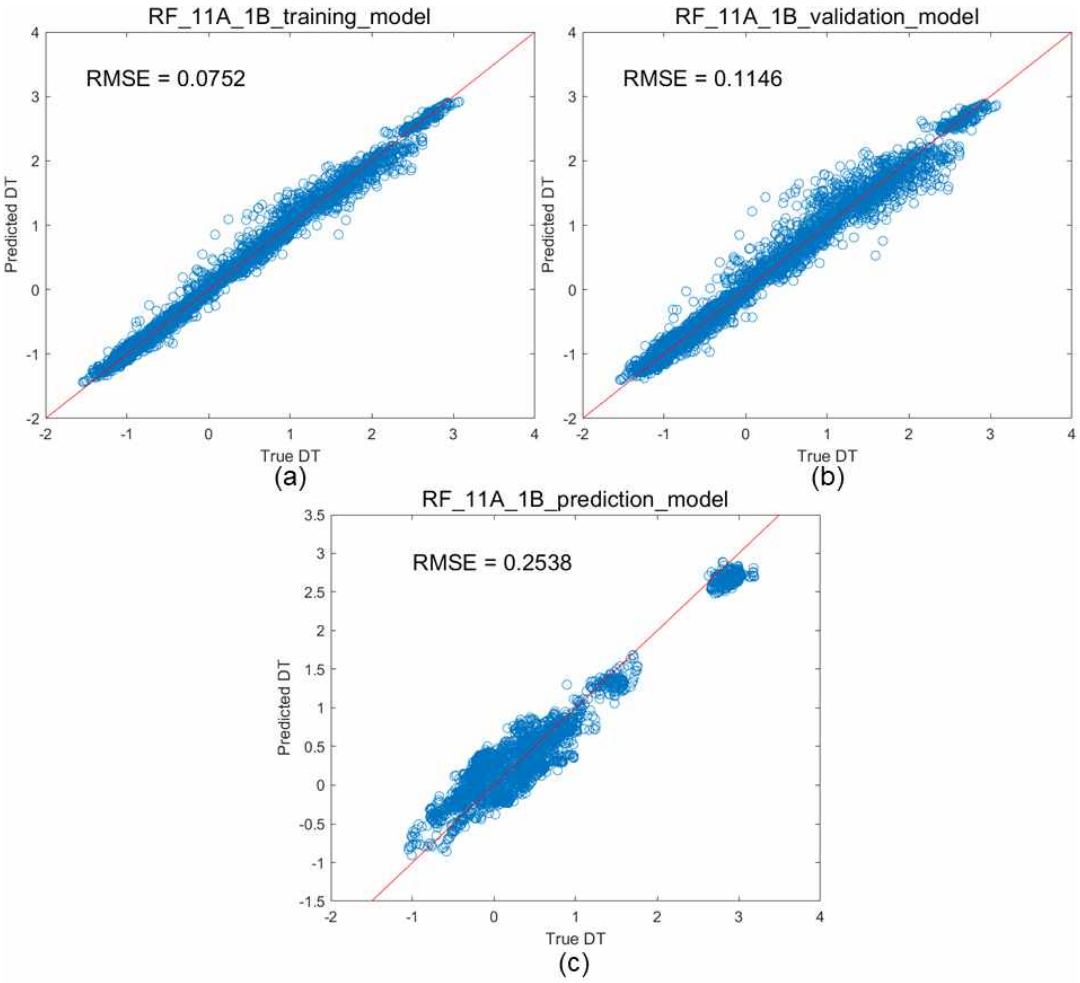


Figure 3.9 (a) Training, (b) validation using 15/9-F-11A data and (c) prediction of DT in 15/9-F-1B.

3) 15/9-F-1A와 15/9-F-11A를 학습자료로 사용한 경우

Figure 3.10은 15/9-F-1A 데이터와 15/9-F-11A 데이터를 학습한 후 15/9-F-1B 데이터의 DT를 예측하는 모델의 결과이다. 모델의 학습 결과를 통해 계산한 RMSE는 0.0851이고, 검증 결과의 RMSE는 0.1301, 15/9-F-1B 데이터의 DT를 예측한 결과의 RMSE는 0.2663이다.

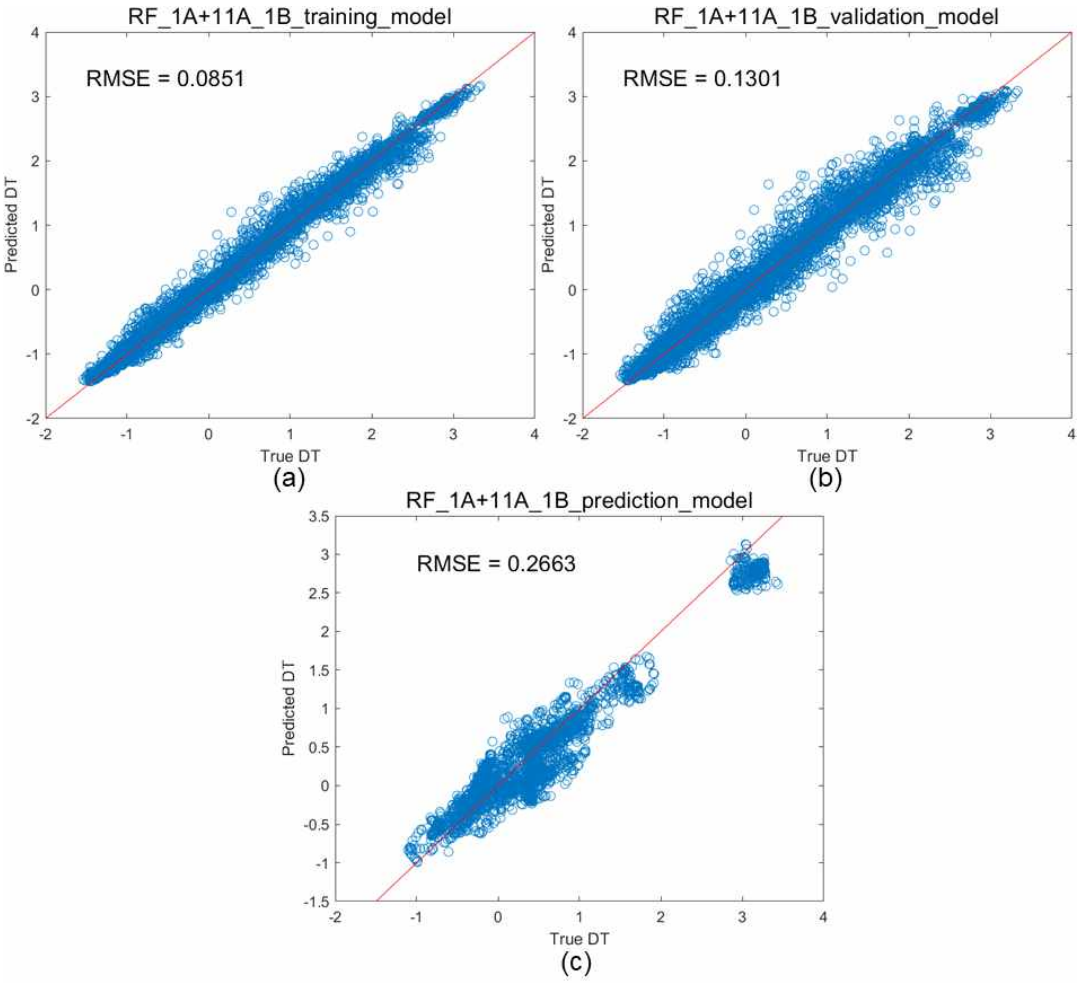


Figure 3.10 (a) Training, (b) validation using 15/9-F-1A+11A data and (c) prediction of DT in 15/9-F-1B.

3. 15/9-F-11A 예측 모델 결과

1) 15/9-F-1A를 학습자료로 사용한 경우

Figure 3.11은 15/9-F-1A 데이터를 학습한 후 15/9-F-11A 데이터의 DT를 예측한 모델의 결과이다. 모델의 학습 결과 RMSE는 0.0872이고, 검증 결과의 RMSE는 0.1345, 15/9-F-11A 데이터의 DT를 예측한 결과의 RMSE는 0.2831이다. 예측 결과 그래프를 보면 예측된 데이터의 끝부분이 실제 값보다 조금 작게 예측된 것을 확인할 수 있었다.

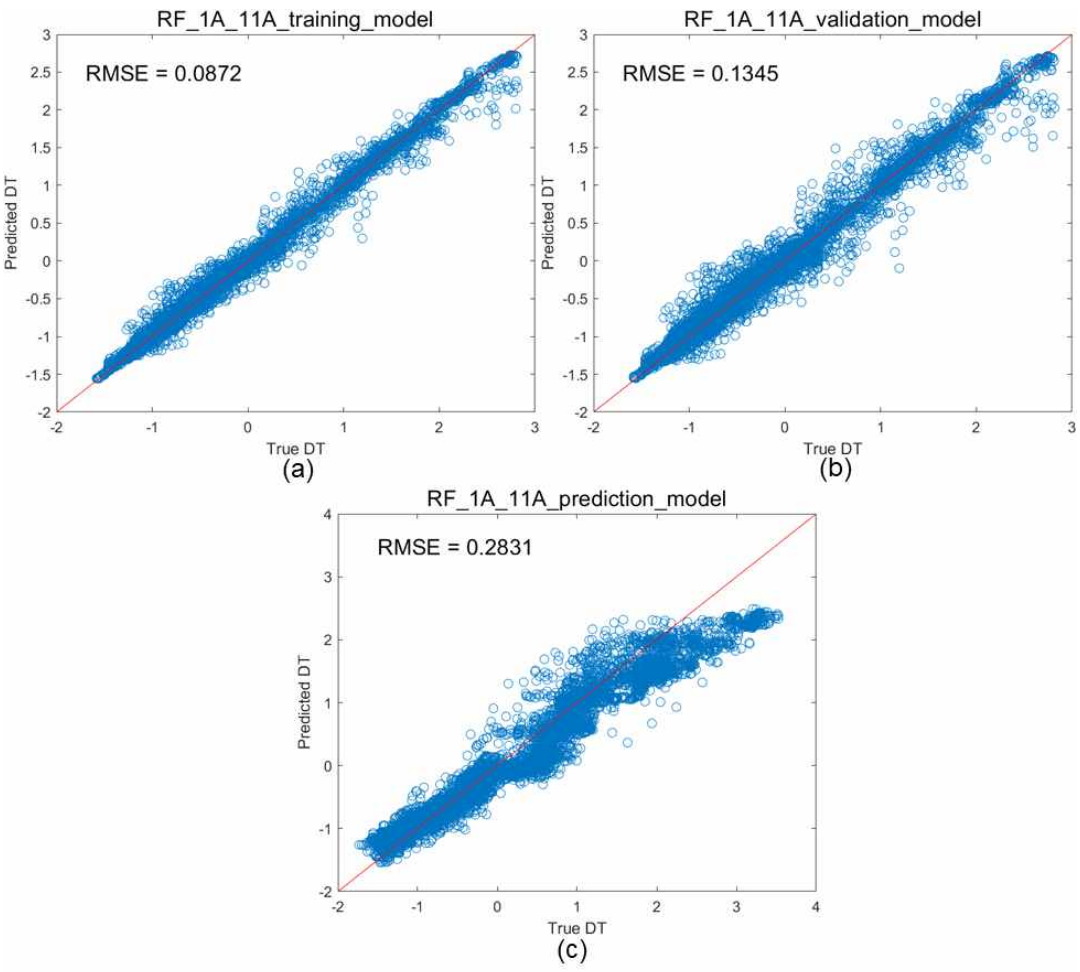


Figure 3.11 (a) Training, (b) validation using 15/9-F-1A data and (c) prediction of DT in 15/9-F-11A.

2) 15/9-F-1B를 학습자료로 사용한 경우

Figure 3.12는 15/9-F-1B 데이터를 학습한 후 15/9-F-11A 데이터의 DT를 예측한 모델의 결과이다. 모델의 학습 결과 RMSE는 0.0792이고, 검증 결과의 RMSE는 0.1224, 15/9-F-11A 데이터의 DT를 예측한 결과의 RMSE는 0.4186이다. 예측 결과 그래프를 보면 예측된 데이터의 앞부분이 실제 값보다 크게 예측되었으며, 뒷부분에도 다수 예측값과 실제값의 매칭이 잘 이루어지지 않음을 확인할 수 있었다.

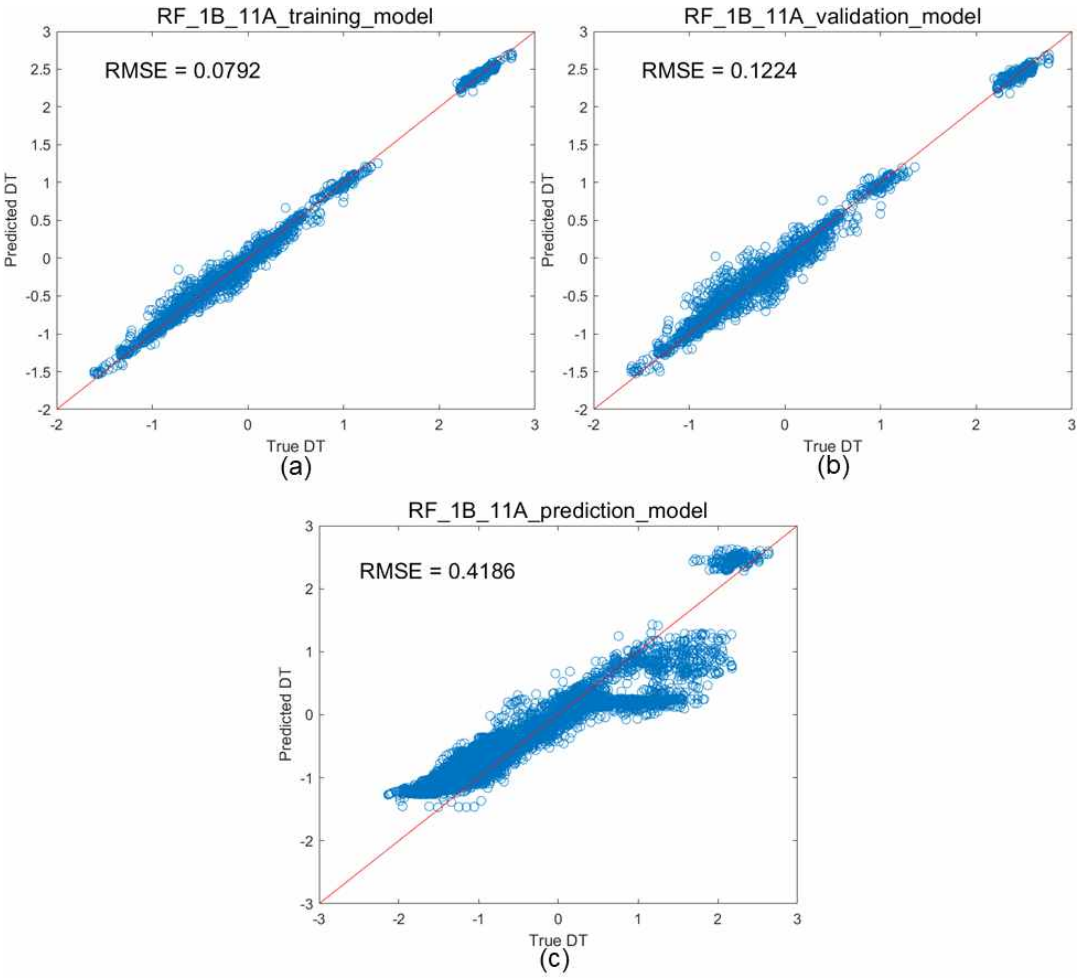


Figure 3.12 (a) Training, (b) validation using 15/9-F-1B data and (c) prediction of DT in 15/9-F-11A.

3)15/9-F-1A와 15/9-F-1B를 학습자료로 사용한 경우

Figure 3.13은 15/9-F-1A 데이터와 15/9-F-1B 데이터를 학습한 15/9-F-11A 데이터의 DT를 예측한 모델의 결과이다. 모델의 학습 결과 RMSE는 0.0872이고, OOB 데이터를 이용한 검증 결과의 RMSE는 0.1344, 15/9-F-11A 데이터의 DT를 예측한 결과의 RMSE는 0.2200이다. 15/9-F-1B 데이터를 단독으로 학습하여 예측한 결과는 좋지 않았지만, 15/9-F-1A 데이터와 함께 사용하여 학습시킨 결과는 좋았다.

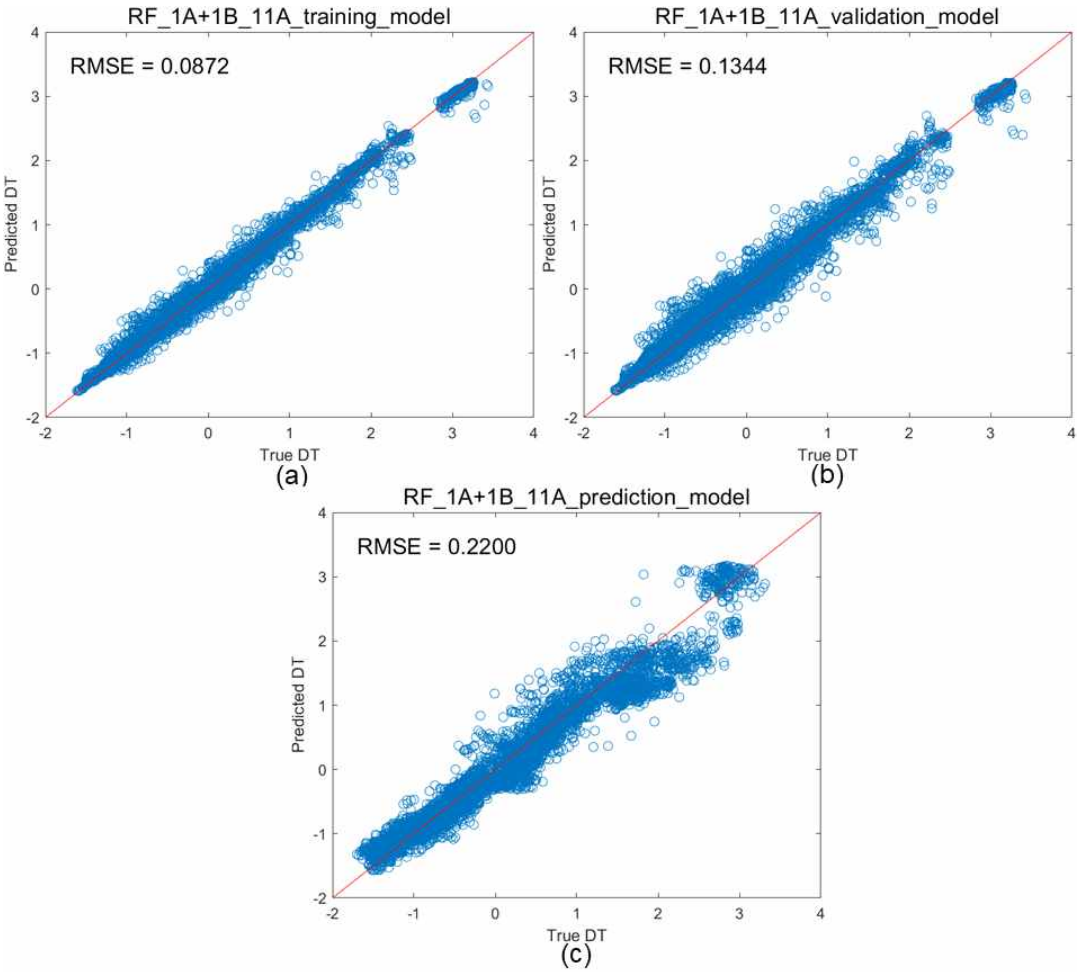


Figure 3.13 (a) Training, (b) validation using 15/9-F-1A+1B data and (c) prediction of DT in 15/9-F-11A.

랜덤 포레스트 기법을 사용하여 다양한 조합을 통해 학습한 모델을 통해 15/9-F-1A, 15/9-F-1B, 15/9-F-11A 데이터의 DT의 예측결과를 Table 3.4에 정리하였다.

Table 3.4 Result of prediction about total models

Model	RMSE	Model	RMSE
RF_1B_1A	0.3973	RF_1B+11A_1A	0.2230
RF_11A_1A	0.2208		
RF_1A_1B	0.5011	RF_1A+11A_1B	0.2663
RF_11A_1B	0.2538		
RF_1A_11A	0.2801	RF_1A+1B_11A	0.2182
RF_1B_11A	0.4186		
Mean	0.3453		0.2358

15/9-F-1A의 DT를 예측한 모델 중 가장 오차가 작은 모델은 15/9-F-11A 데이터를 학습한 모델이며, 15/9-F-1B와 15/9-F-11A 데이터를 학습한 모델도 좋은 성능을 보였다.

15/9-F-1B의 DT를 예측한 모델 중 가장 오차가 작은 모델은 15/9-F-11A 데이터를 학습한 모델이며, 15/9-F-1A와 15/9-F-11A 데이터를 학습한 모델도 좋은 성능을 보였다.

15/9-F-11A의 DT를 예측한 모델 중 가장 오차가 작은 모델은 15/9-F-1A와 15/9-F-1B 데이터를 학습한 모델이었다.

한 개의 시추공 데이터를 학습하여 예측한 결과의 평균 RMSE는 0.3453이고, 두 개의 데이터를 학습하여 예측한 결과의 평균 RMSE는 0.2358이다. 두 개의 시추공 데이터를 함께 사용하여 학습모델을 구축했을 때 평균 RMSE가 31.71% 더 낮았다. 두 개의 시추공 데이터를 학습하여 예측한 결과의 평균 오차가 더 좋은 이유를 분석해보면, 하나의 데이터를 이용할 때보다 다양한 입력 데이터의 분포에 대하여 학습하기 때문에 대체적으로 좋은 성능을 보이는 것으로 판단되었다. 이와 같은 결과를 바탕으로 15/9-F-1A, 15/9-F-1B, 15/9-F-11A를 모두 학습한 모델을 이용하면

다양한 입력 데이터의 분포에 대해 학습하므로 음파 검층이 미측정된 15/9-F-1C, 15/9-F-11B 데이터의 DT를 예측할 때 결과가 가장 좋을 것으로 예상된다.

4절에서는 SVDD 기법을 이용하여 전체 예측 결과를 신뢰도에 따라 영역을 구분하는 방법을 제안하였다.

제4절 SVDD 기법을 이용한 예측 결과 분석

SVDD는 주로 이상치를 탐지하기 위해 사용되는 기법으로 탐색 공간(feature space)에서 대부분의 정상 데이터를 둘러싸는 가장 작은 구(hypersphere)의 경계를 찾는 것을 목적으로 한다. 따라서 해당 기법을 이용하여 학습데이터를 기반으로 경계를 설정하여 경계 안에 위치한 데이터들의 예측 결과와 경계 밖에 위치한 데이터들의 예측 결과를 비교할 수 있다.

이 절에서는 제 3절에서 생성한 랜덤 포레스트 예측 모델의 결과에 대해 SVDD 기법을 적용하여 예측 신뢰도 분석을 실시하였다. 이를 위해 SVDD의 하이퍼파라메터 중 하나인 g값을 변화시켜 민감도 분석을 실시하였다. 여기서 g값은 커널 너비의 역수를 말하며, 가우시안 커널인 식 (3)의 s와의 관계는 식(7)과 같다.

$$g = \frac{1}{s} \tag{7}$$

g 값의 변화에 따라서 기준이 되는 학습 데이터의 중심으로부터 구의 경계 사이의 거리가 달라지며 g값이 클수록 학습 데이터에 가깝게 경계가 설정된다. 학습된 영역 내에 포함된 데이터를 예측하는 경우 오차가 작을 것으로 예상되는 반면, 학습 데이터의 경계 외부에 위치한 데이터를 예측하는 경우에는 학습된 공간 범위를 넘어서 값을 추정하는 외삽법(extrapolation)에 해당하므로 오차가 클 것으로 예상된다.

이를 확인하기 위하여 g값을 1, 3, 5로 점점 증가시켜 학습 데이터를 기준으로 경계를 설정하였으며, 생성된 경계를 기준으로 데이터를 분리하여 예측 결과를 비교하였다. 경계 내부에 위치한 데이터는 편의상 'inBND' 데이터, 외부에 위치한 데이터는 'outBND' 데이터로 명명하여 용어를 간단히 하였다.

1. 15/9-F-1A 예측 모델 결과

1) 15/9-F-1B를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.14는 (a)와 (b)는 15/9-F-1B를 학습 데이터로 사용하고 SVDD의 g=1로 설정하였을 때 생성되는 경계를 기준으로 15/9-F-1A의 입력자료를 내부(inBND) 및 외부(outBND) 데이터로 나눈 후 예측한 DT 결과이다. inBND와 outBND 예측 결과가 확연히 다른 모습을 볼 수 있다. 이것은 모델의 학습 데이터로 사용된 15/9-F-1B 데이터로 설명하기 어려운 15/9-F-1A 데이터가 경계를 기준으로 비교적 잘 나눠졌음을 의미하며, 15/9-F-1B와 15/9-F-1A 데이터 간의 유사도가 낮다는 것을 알 수 있다. 이를 통해 앞서 15/9-F-1B를 학습데이터로 사용하여 만든 모델이 15/9-F-1A의 데이터에 대한 예측 성능이 좋지 않은 이유를 설명할 수 있다.

15/9-F-1A와 15/9-F-1B의 입력 데이터의 유사도를 알기 위해 입력변수에 대한 2차원 맵핑을 수행하였다. 입력 데이터는 5차원 변수이며 데이터의 차원 축소에 사용되는 알고리즘인 tSNE(t-distributed stochastic neighbor embedding)을 이용하여 데이터를 2차원으로 가시화하였다. tSNE는 비슷한 데이터는 근접하게, 다른 데이터는 멀리 떨어진 지점으로 나타낸다(Maaten 등, 2008). inBND 영역을 나타낸 Figure 3.14(c)는 두 데이터가 중첩되거나 일부 근접하지 않은 데이터까지 포함하는 것을 알 수 있으며, outBND영역을 나타낸 Figure 3.14(d)는 15/9-F-1A 데이터와 15/9-F-1B가 적절히 분리된 것으로 나타났다.

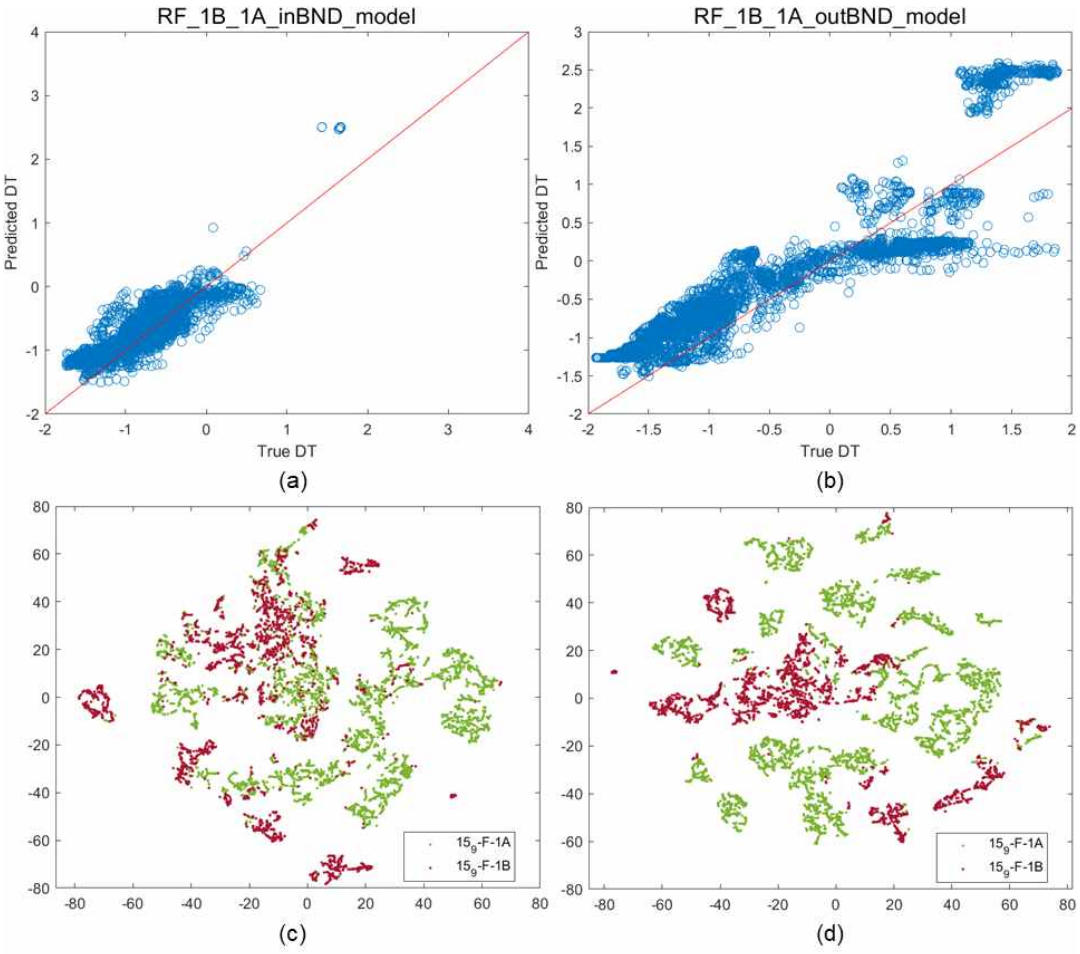


Figure 3.14 Prediction result of RF_1B_1A_inBND and RF_1B_1A_outBND model for g=1

데이터를 3차원으로 맵핑하여 비교하기 위해 또 다른 차원축소 방법인 PCA를 사용하여 두 데이터의 입력변수를 Figure 3.15와 같이 나타내었다. Figure 3.15 (a)는 데이터를 분리하기 전의 15/9-F-1B와 15/9-F-1A 분포 형태이며, Figure 3.15(b)는 inBND를 표시한 것으로 15/9-F-1B의 학습데이터가 분포한 공간 내부 및 경계에 15/9-F-1A 데이터가 분포되어 있는 것을 확인할 수 있다. Figure 3.15 (c)는 outBND 데이터 분포로 inBND과 확실한 차이를 보인다.

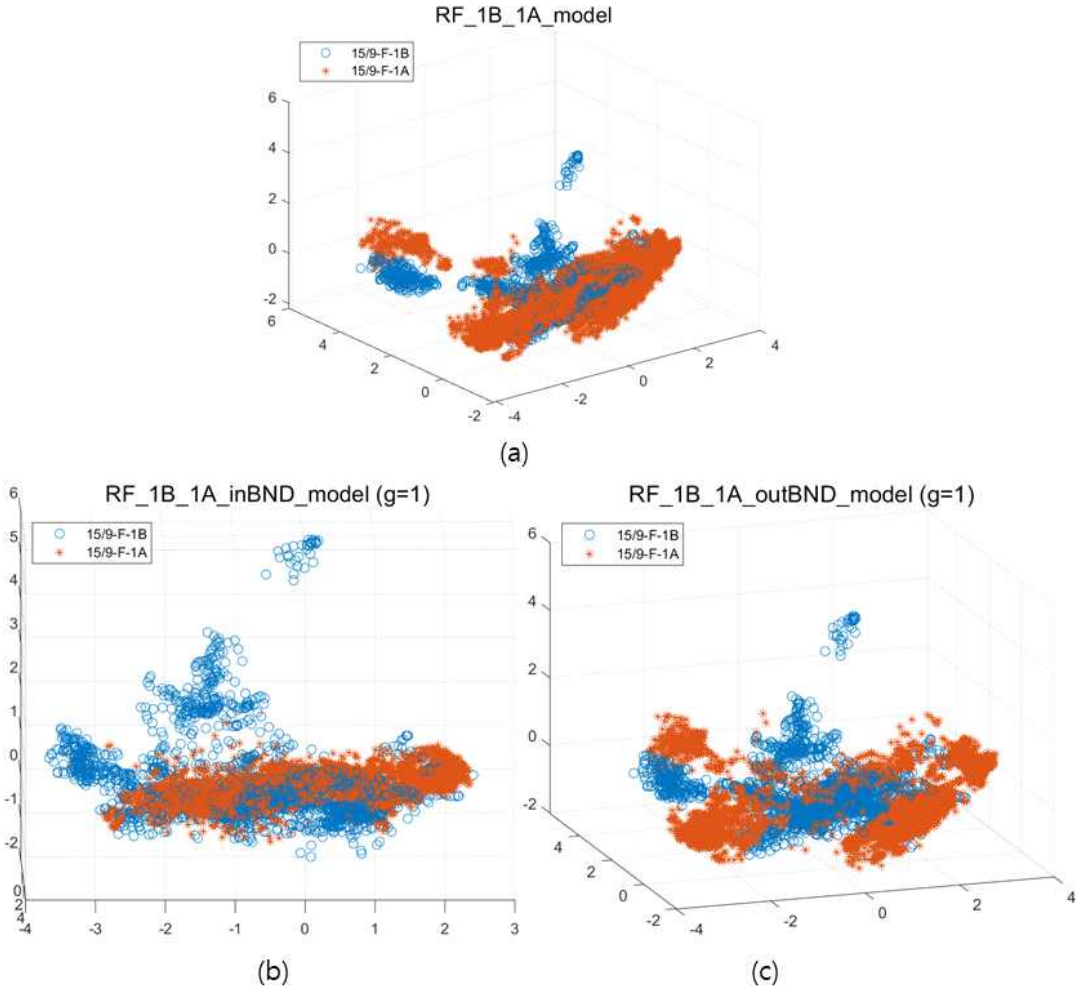


Figure 3.15 Result of PCA of RF_1B_1A_model, inBND_model and outBND_model for g=1.

나. g=3인 경우

Figure 3.16는 g=3로 설정하였을 때 15/9-F-1A의 DT를 예측한 결과이다. Figure 3.16(a)와 Figure 3.14(a)를 비교하면 inBND 데이터의 개수가 감소한 것을 확인할 수 있으며, inBND 위치를 나타내는 Figure 3.16(c)는 Figure 3.14(c)와 비교하여 경계에 근접한 데이터로 줄어든 것을 볼 수 있다.

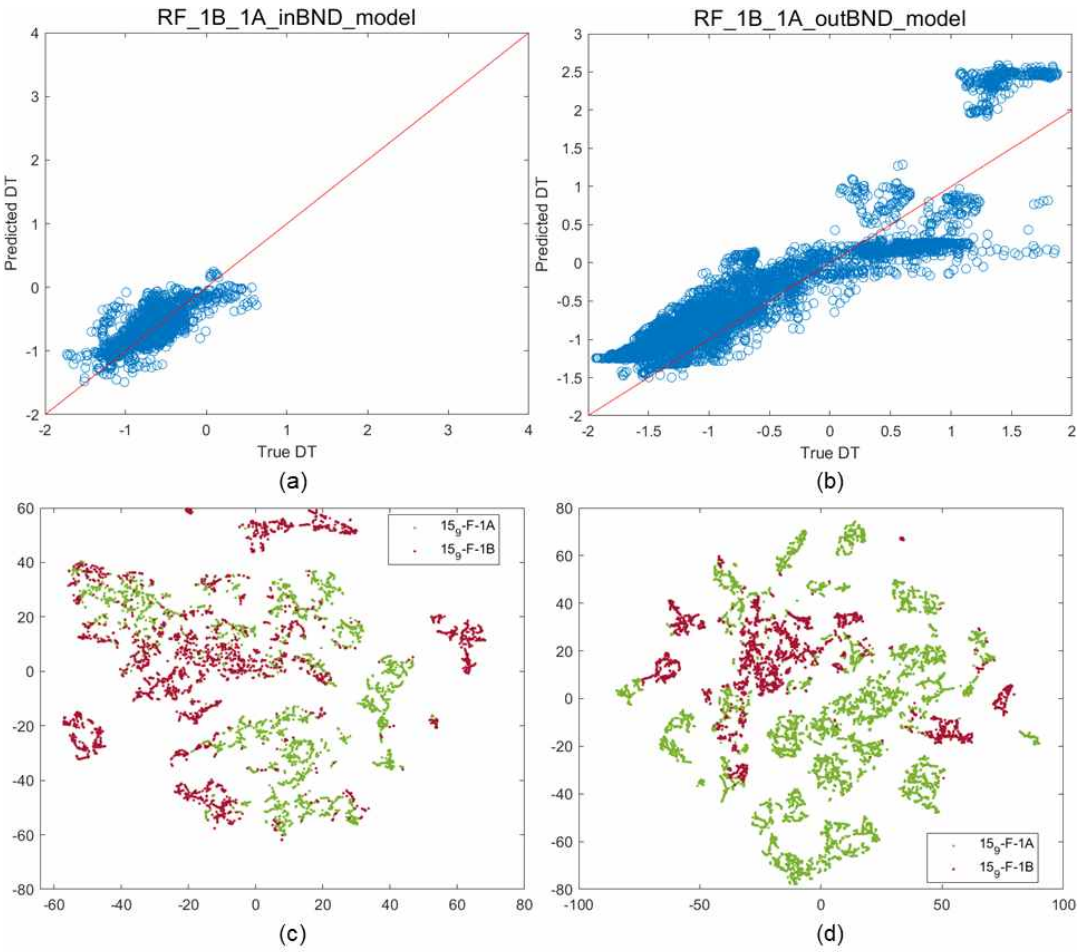


Figure 3.16 Prediction result of RF_1B_1A_inBND and RF_1B_1A_outBND model for g=3.

다. g=5인 경우

Figure 3.17은 g=5로 설정하였을 때 15/9-F-1A의 DT를 예측한 결과이다. Figure 3.17(a) 그래프에서 확인할 수 있듯 아주 적은 양의 데이터만인 inBND영역에 남아있고 대부분의 데이터가 Figure 3.17(b) 그래프와 같이 outBND에 존재하는 것을 확인할 수 있다. inBND영역을 나타내는 Figure 3.17(c)는 Figure 3.16(c)와 비교하여 아주 적은 양인 것을 볼 수 있고 대부분의 데이터가 outBND 영역에 존재하는 것을 확인할 수 있다.

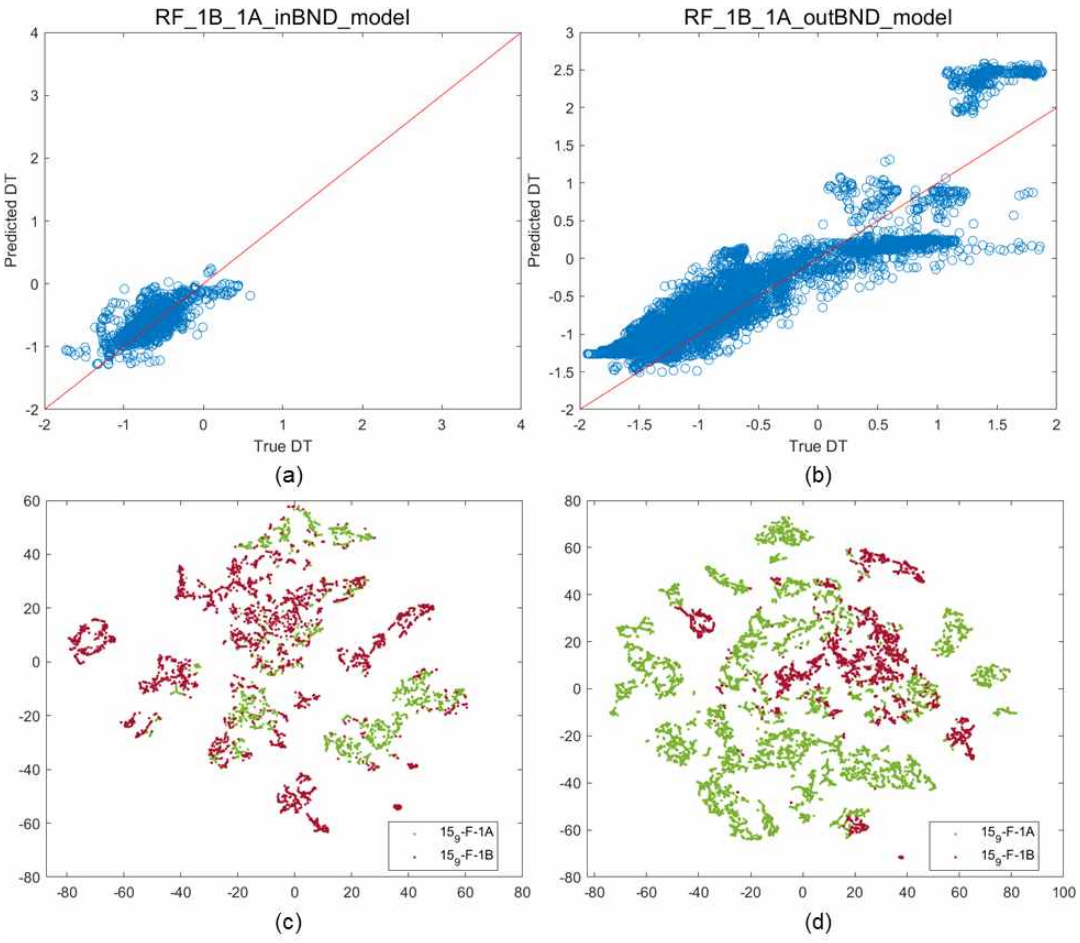


Figure 3.17 Prediction result of RF_1B_1A_inBND and RF_1B_1A_outBND model for g=5.

Figure 3.18은 15/9-F-1B 데이터를 학습하여 15/9-F-1A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.3973이다.

inBND RMSE는 g=1일 때 0.2491, g=3일 때 0.2033, g=5일 때 0.2004로 점점 오차가 감소하는 것으로 나타났으며, 전체 데이터를 사용하여 학습한 모델보다 작은 오차로 나타났다. 또한 15/9-F-1B 데이터와 연관성이 가장 높은 데이터인 g=5일 때의 오차가 가장 작은 것을 알 수 있다. 반면, outBND RMSE는 g=1일 때 0.4933, g=3일 때 0.4488, g=5일 때 0.4252로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다. 이를 통해 SVDD를 사용하여 학습데이터의 경계를 설정하고 그 경계 범위 내에 위치한 데이터들을 사용했을 때의 예측 성능이 뛰어남을 알 수 있다.

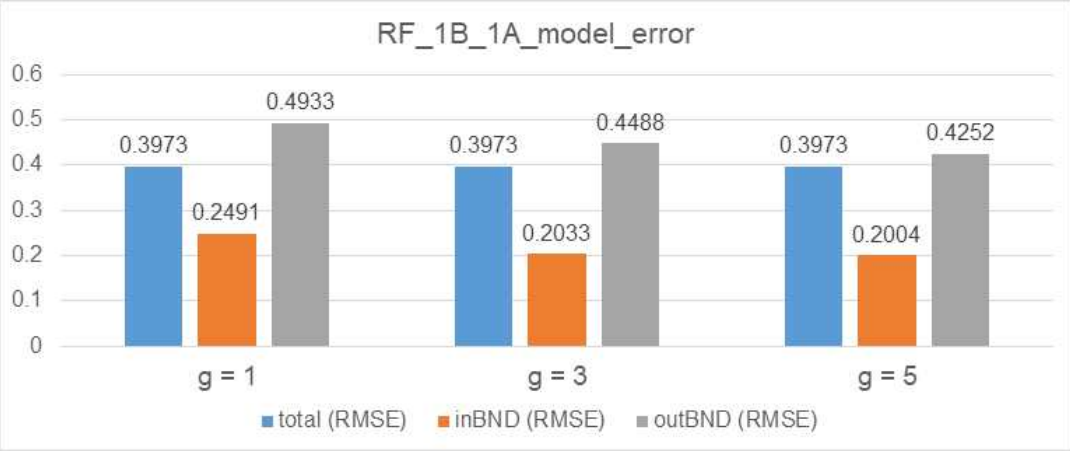


Figure 3.18 Comparison result of RF_1B_1A_model for g value change

Figure 3.19는 15/9-F-1B 데이터를 학습하여 15/9-F-1A를 예측하는 모델을 g=1, g=5인 경우 inBND와 outBND 데이터의 오차(= 참값 - 예측값)를 히스토그램으로 나타낸 그래프이다. inBND 데이터에 오차가 0에 가까운 빈도수가 outBND 데이터보다 많으며, g=5인 경우 inBND 데이터에 있던 0에 가까운 오차가 outBND 데이터에 포함된 것을 확인하였다.

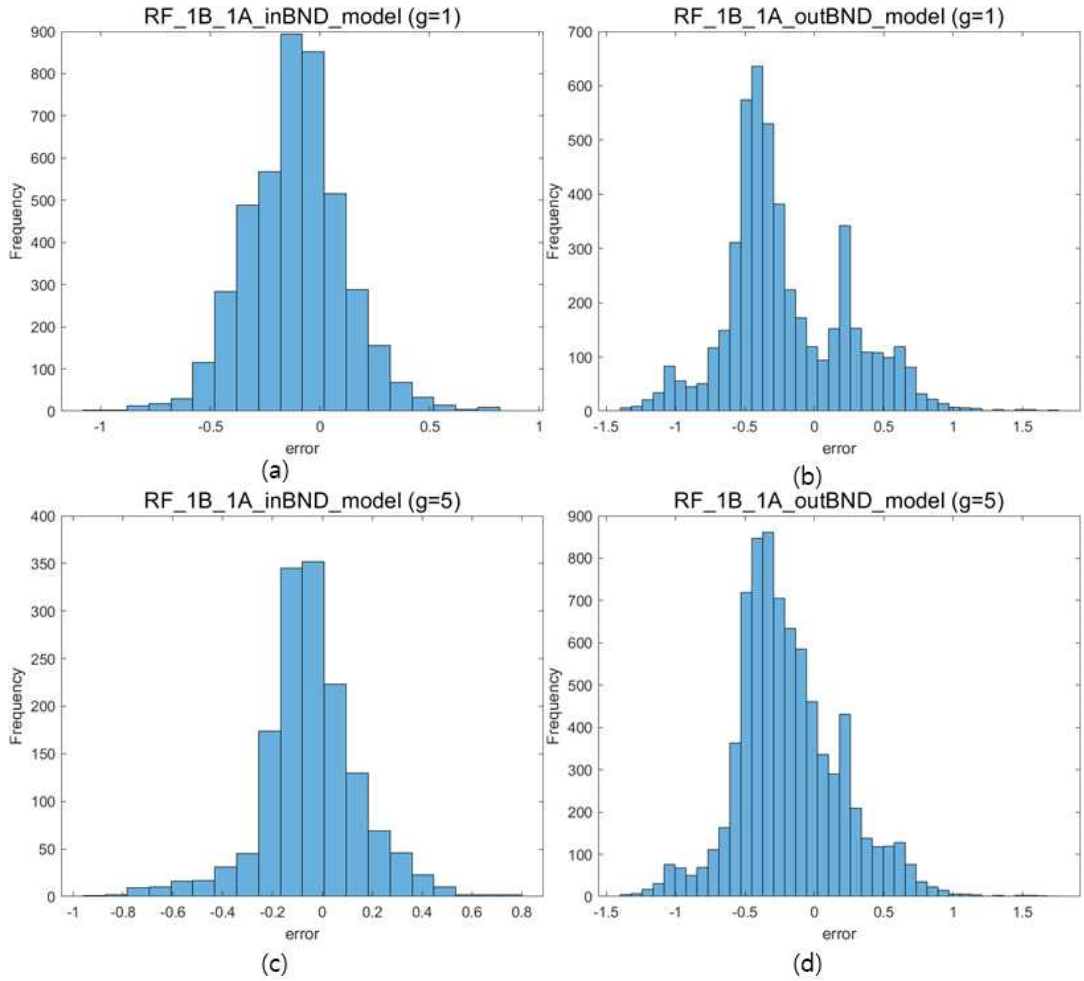


Figure 3.19 Histogram of RF_1B_1A_inBND_model error and RF_1B_1A_outBND_model error for g value change.

Table 3.5는 15/9-F-1B 데이터를 학습하여 15/9-F-1A의 DT를 예측한 모델을 g값에 따른 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각각 47.17%, 27.23%, 16.35%로 점점 감소하였으며, 경계 외부에 해당하는 데이터의 비율은 각각 52.83%, 72.77%, 83.65%로 증가하였다. 전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 37.30%, g=3일 때, 48.83%, g=5일 때, 49.56% 감소하였으며, outBND RMSE는 g=1일 때, 24.16%, g=3일 때, 12.96% , g=5일 때, 7.02% 증가하였다.

전체적으로 해당 모델에서의 g값 변화에 따른 결과를 비교했을 때, g값이 증가함에 따라 inBND 데이터 오차는 전체 오차보다 점점 감소하였다. outBND 데이터의 오차 역시 g값이 증가함에 따라 감소하는 경향을 보이는데, 그 이유는 g값이 커짐에 따라 경계가 축소되고 inBND 데이터의 좋은 영역에 포함되어 있던 데이터가 outBND 데이터로 이동하면서 outBND 데이터의 오차를 낮추는 효과를 보이는 것으로 분석되었다.

Table 3.5 Comparison result of RF_1B_1A_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	47.17 %	27.23 %	16.35 %
ratio of data out of boundary	52.83 %	72.77 %	83.65 %
(1)total RMSE	0.3973		
(2)inBND RMSE	0.2491	0.2033	0.2004
(3)outBND RMSE	0.4933	0.4488	0.4252
relative difference between (1) and (2)	-37.30 %	-48.83 %	-49.56 %
relative difference between (1) and (3)	24.16 %	12.96 %	7.02 %

Figure 3.20은 15/9-F-1A 데이터의 NPHI, RHOB, GR, RT, PEF와 예측된 DT를 깊이에 따라 그래프로 나타낸 것이다. 실제 DT는 파란색 선이고 예측된 DT는 주황색 선으로 나타냈다. Figure 3.20(a)의 색칠된 영역은 g=1인 경우 전체 데이터의 47.17%에 해당하는 inBND 데이터의 예측 결과에 해당하며, 해당 영역의 오차는 전체 오차보다 37.30% 낮다. Figure 3.20(b)의 색칠된 영역은 g=3인 경우 전체 데이터의 27.33%에 해당하는 inBND 데이터의 예측 결과이다. 해당 영역의 오차는 전체 오차보다 48.83% 낮다. Figure 3.20(c)의 색칠된 영역은 g=5인 경우 전체 데이터의 16.35%에 해당하는 inBND 데이터의 예측 결과이며, 해당 영역의 오차는 전체 오차보다 49.56% 낮다. Figure 3.20(d)는 (a),(b),(c)를 중첩하여 나타낸 결과이다.

Figure 3.20의 A, B, C, D영역은 서로 대비되는 뚜렷한 특징을 가지는 영역을 분류한 것이다. A영역은 g=1, 3, 5에 관계없이 대부분 outBND에 속하는 것을 표시한 것이며, B영역은 g=1일 때, 대부분 inBND 데이터로 해석되지만 g=3, 5일 때는 대부분 outBND에 속하는 것을 나타낸 것이다. 그리고 C영역은 g=1일 때, 대부분 inBND 데이터이지만 g=3, 5일 때는 간헐적으로 outBND로 분류되는 것이며, D영역은 g=1, 3, 5일 때 모두 inBND로 분류 영역을 표시한 것이다.

A영역은 학습 데이터가 분포하는 공간에서 벗어나 다소 먼 거리에 위치하는 것으로 해석되며, B영역은 경계에 가까운 바깥쪽에 분포할 것으로 예상된다. 그리고 C영역은 학습 데이터 경계 주변에 혼재되어 있을 것으로 추정되며, D영역은 학습 데이터 공간 내에 분포되어 있을 것으로 해석된다. 이와 같이 g=1, 3, 5로 변화시켜 가면서 경계영역 내에 존재하는 데이터를 분석하면 예측결과에 대해 높은 신뢰도를 보이는 구간을 선정할 수 있다. Figure 3.20과 같이 C와 D영역에 해당하는 구간이 신뢰도가 높은 구간으로 추정되었고 실제값과 비교하여 타당함을 확인하였다.

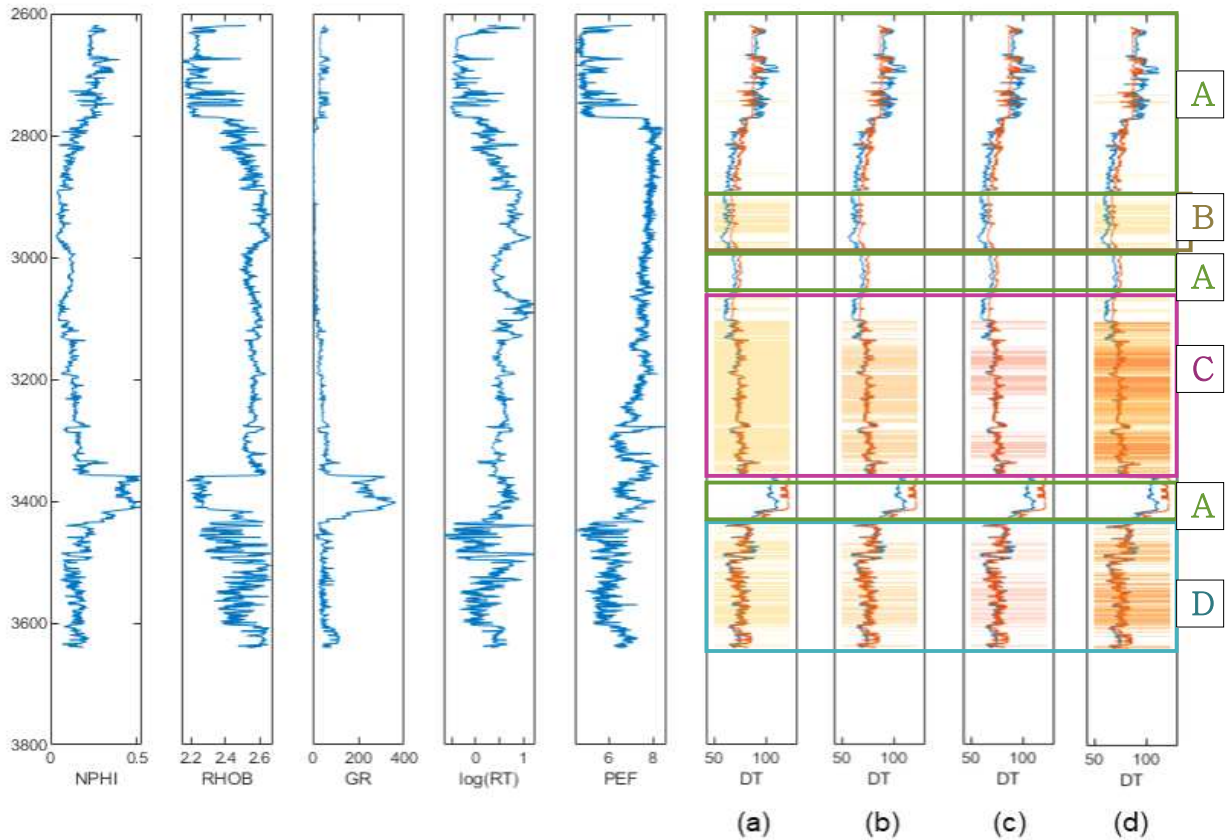


Figure 3.20 Well log of input data and predicted DT of RF_1B_1A_model for g value change.

2) 15/9-F-11A를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.21(a)와 (b)는 15/9-F-11A 데이터를 기준으로 SVDD를 사용해 g=1로 설정하였을 때 생성되는 경계를 기준으로 15/9-F-1A의 DT를 예측한 결과이다. outBND 모델은 일부 데이터가 실제보다 큰 값으로 예측된 결과를 보였다. Figure 3.21(c)에서 inBND 데이터와 학습데이터의 거리는 Figure 3.21(d)의 outBND 데이터의 거리보다 가깝게 분석되었다.

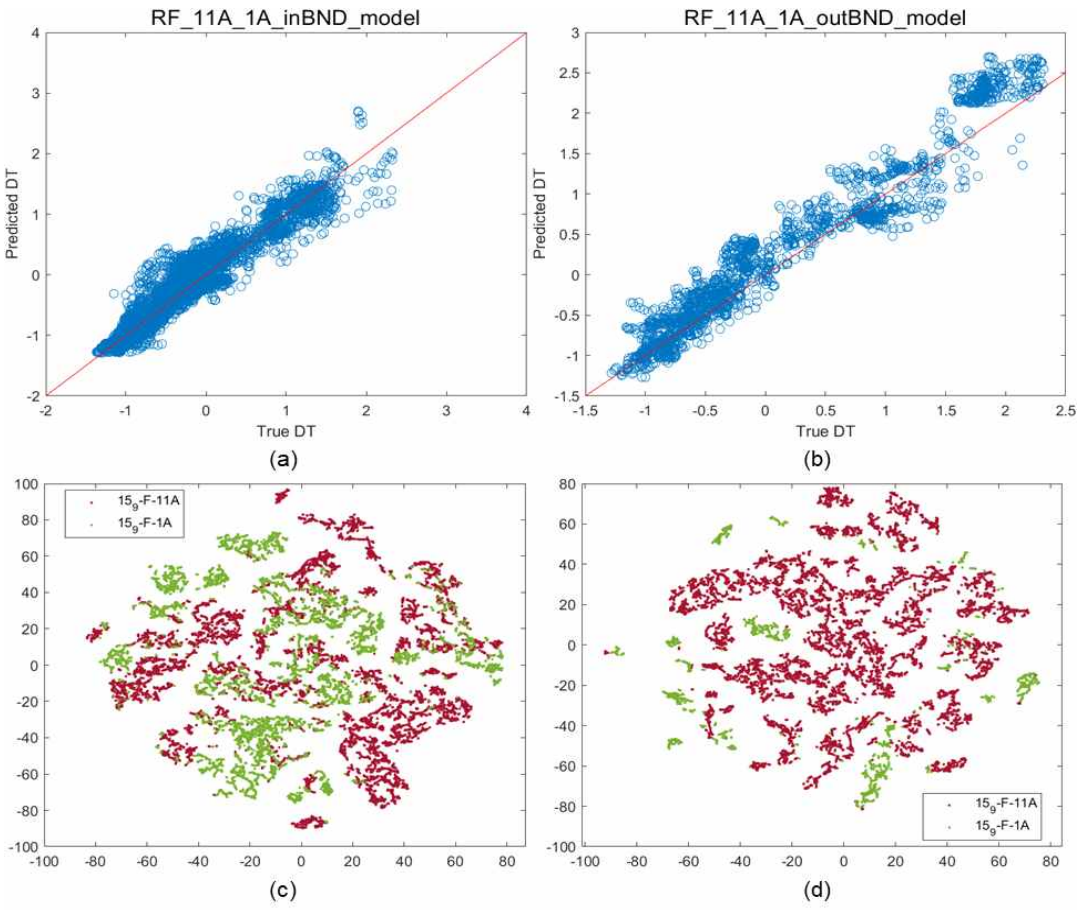


Figure 3.21 Prediction result of 15/9-F-1A using 15/9-F-11A training model for g=1.

나. g=5인 경우

Figure 3.22(a)와 (b)는 15/9-F-11A 데이터를 기준으로 SVDD를 사용해 g=5로 설정하였을 때 생성되는 경계를 기준으로 15/9-F-1A의 DT를 예측한 결과이다.

Figure 3.21(a), (b)와 비교했을 때, g=5인 경우 inBND의 데이터가 outBND로 이동한 것을 볼 수 있다. 또한 Figure 3.22(c), (d)를 통해 inBND 영역에 가까운 데이터만 남아 있는 것을 확인할 수 있었다.

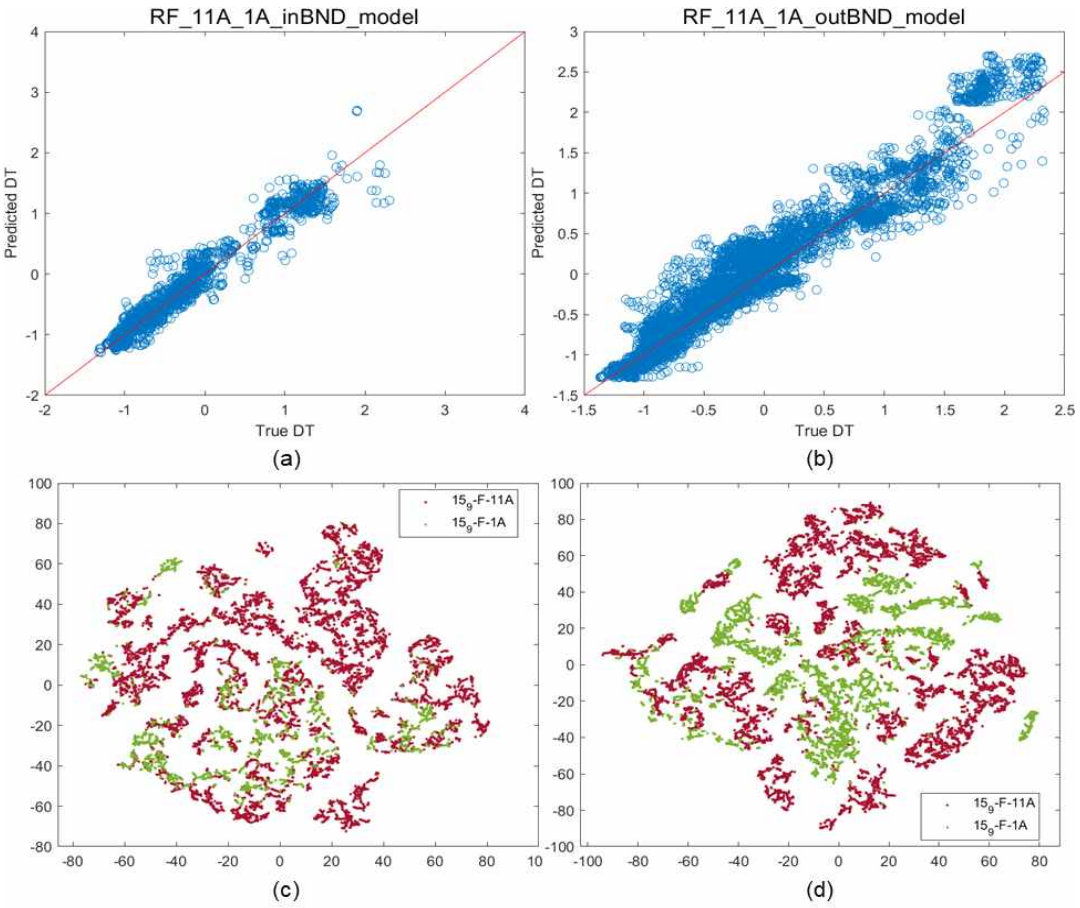


Figure 3.22 Prediction result of 15/9-F-1A using 15/9-F-11A training model for g=5.

Figure 3.23은 15/9-F-11A 데이터를 학습하여 15/9-F-1A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.2208이다.

inBND RMSE는 g=1일 때 0.1944, g=3일 때 0.1820, g=5일 때 0.1654로 점점 오차가 감소하는 것으로 나타났으며, 전체 데이터를 가지고 학습한 모델보다 작은 오차로 나타났으며, 15/9-F-1A 데이터와 연관성이 가장 높은 데이터인 g=5일 때의 오차가 가장 작게 나타났다. 반면, outBND RMSE는 g=1일 때 0.2966, g=3일 때 0.2454, g=5일 때 0.2383으로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다. 이를 통해 SVDD를 사용하여 학습데이터의 경계를 설정하고 그 경계 범위 내에 위치한 데이터들을 사용했을 때의 예측 성능이 뛰어남을 알 수 있다.

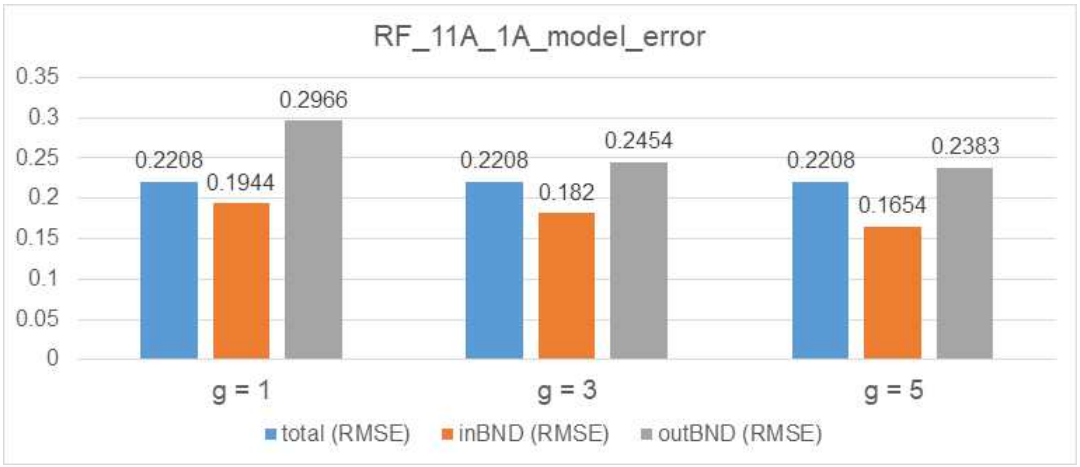


Figure 3.23 Comparison result of RF_11A_1A_model for g value change.

Table 3.6은 15/9-F-11A 데이터를 학습하여 15/9-F-1A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각 각 78.19%, 42.30%, 27.28%이며, 경계 외부에 해당하는 데이터의 비율은 21.81%, 57.70%, 72.72%로 증가하였다.

전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 11.96%, g=3일 때, 17.57%, g=5일 때, 25.09% 감소하였으며, outBND RMSE는 g=1일 때, 34.33%, g=3일 때, 11.14%, g=5일 때, 7.93% 증가하였다.

Table 3.6 Relative difference result of RF_11A_1A_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	78.19%	42.30%	27.28%
ratio of data out of boundary	21.81%	57.70%	72.72%
(1)total RMSE	0.2208		
(2)inBND RMSE	0.1944	0.1820	0.1654
(3)outBND RMSE	0.2966	0.2454	0.2383
relative difference between (1) and (2)	-11.96%	-17.57%	-25.09%
relative difference between (1) and (3)	34.33%	11.14%	7.93%

Figure 3.24는 15/9-F-11A를 학습하여 15/9-F-1A의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간 중 절반 이상이 C, D 영역으로 구분되었으며, A, B 영역은 예측 구간의 끝 부분에 나타났다.

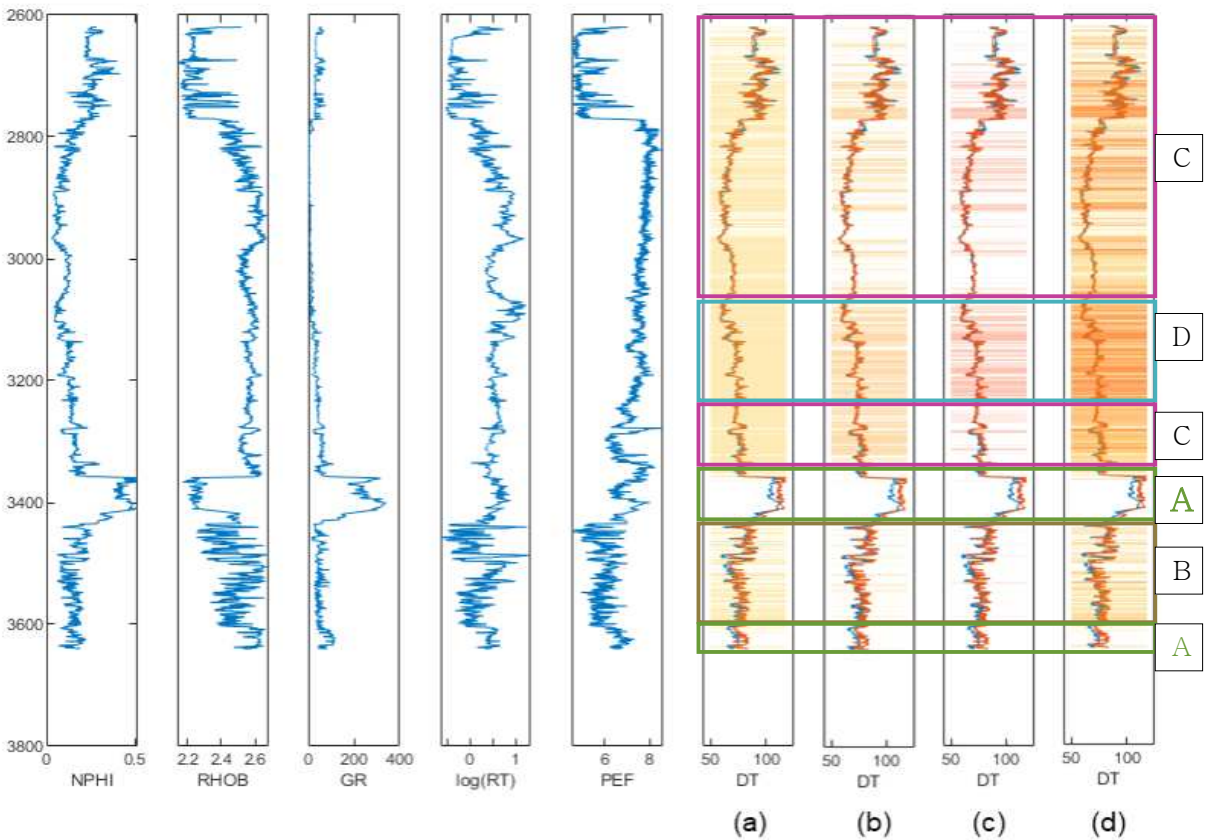


Figure 3.24 Well log of input data and predicted DT of RF_11A_1A_model for g value change.

3) 15/9-F-1B와 15/9-F-11A를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.25(a)와 (b)는 15/9-F-1B와 15/9-F-11A 데이터를 기준으로 SVDD를 사용해 g=1로 설정하였을 때 생성되는 경계를 기준으로 15/9-F-1A의 DT를 예측한 결과이다. 실제값보다 예측이 크게 된 데이터가 outBND 데이터에 포함되었다. Figure 3.25(c)에서 학습데이터와 거리가 먼 데이터가 inBND 모델에 일부 포함되어있는 것을 확인하였다.

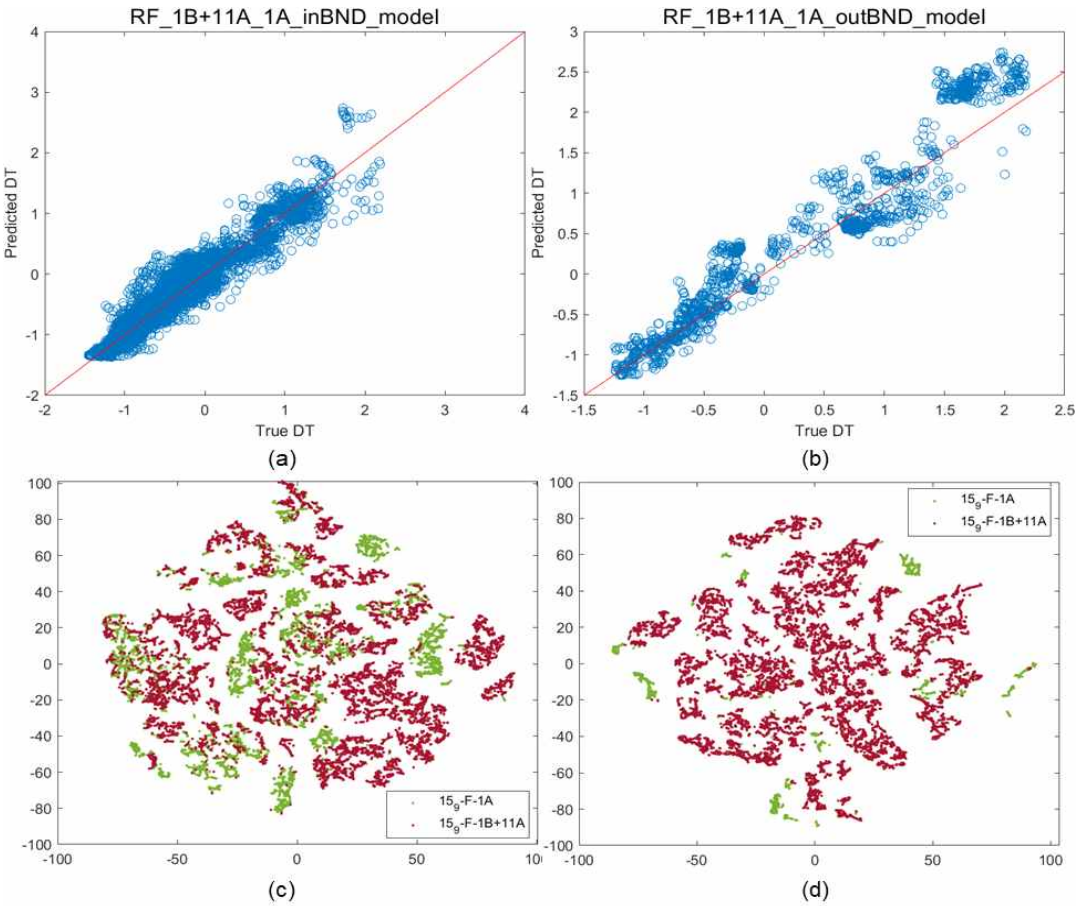


Figure 3.25 Prediction result of 15/9-F-1A using 15/9-F-1B and 15/9-F-11A training model for g=1.

나. g=5인 경우

Figure 3.26(a)와 (b)는 15/9-F-1B와 15/9-F-11A 데이터를 기준으로 SVDD를 사용해 g=5로 설정하였을 때 생성되는 경계를 기준으로 15/9-F-1A의 DT를 예측한 결과이다. Figure 3.25(a), (b)와 비교했을 때, g가 1에서 5로 증가함에 따라 실제값보다 예측이 크게 된 데이터가 outBND 데이터로 다수 이동한 것을 확인하였다. Figure 3.26(c)의 inBND와 학습데이터의 거리가 Figure 3.25(c)의 거리보다 가까운 것으로 분석되었다.

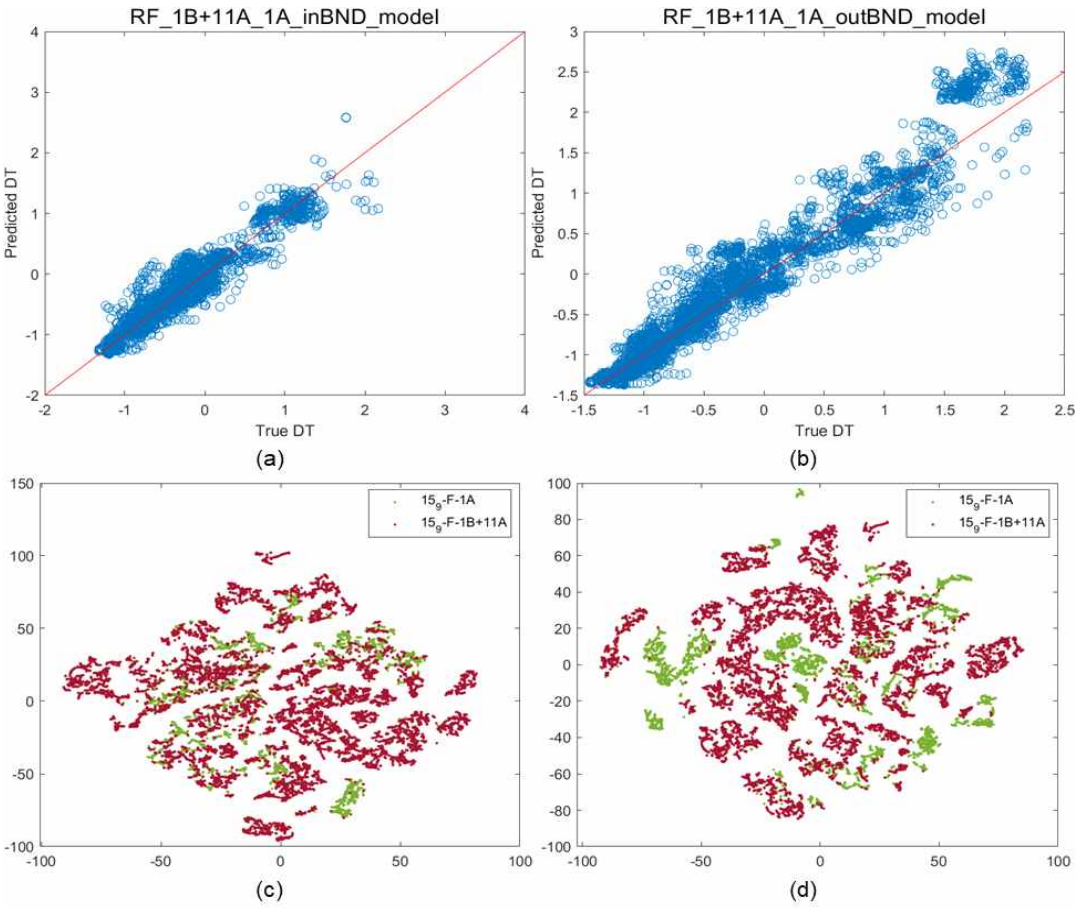


Figure 3.26 Prediction result of 15/9-F-1A using 15/9-F-1B and 15/9-F-11A training model for g=5.

Figure 3.27은 15/9-F-1B와 15/9-F-11A 데이터를 학습하여 15/9-F-1A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.2230이다.

inBND RMSE는 g=1일 때 0.1847, g=3일 때 0.1830, g=5일 때 0.1777로 점점 오차가 감소하는 것으로 나타났으며, 전체 데이터를 가지고 학습한 모델보다 작은 오차로 나타났으며, g=5일 때의 오차가 가장 작았다. 반면, outBND RMSE는 g=1일 때 0.3684, g=3일 때 0.2735, g=5일 때 0.2533으로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다.

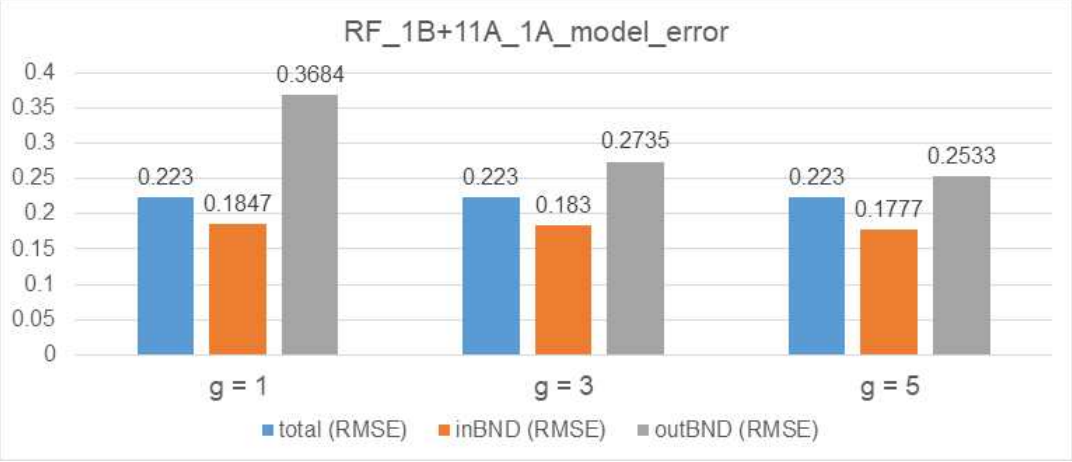


Figure 3.27 Comparison result of RF_1B+11A_1A_model for g value change.

Table 3.7은 15/9-F-1B와 15/9-F-11A 데이터를 학습하여 15/9-F-1A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각각 84.62%, 60.65%, 44.29%이며, 경계 외부에 해당하는 데이터의 비율은 15.38%, 39.35%, 55.71%로 증가하였다.

전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 17.17%, g=3일 때, 17.94%, g=5일 때, 20.31% 감소하였으며, outBND RMSE는 g=1일 때, 65.20%, g=3일 때, 22.65%, g=5일 때, 13.59% 증가하였다.

Table 3.7 Relative difference result of RF_1B+11A_1A_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	84.62 %	60.65 %	44.29 %
ratio of data out of boundary	15.38 %	39.35 %	55.71 %
(1)total RMSE	0.223		
(2)inBND RMSE	0.1847	0.1830	0.1777
(3)outBND RMSE	0.3684	0.2735	0.2533
relative difference between (1) and (2)	-17.17 %	-17.94 %	-20.31 %
relative difference between (1) and (3)	65.20 %	22.65 %	13.59 %

Figure 3.28은 15/9-F-1B와 15/9-F-11A 데이터를 학습한 모델로 15/9-F-1A의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간 중 일부분이 A 영역에 해당하지만, 해당 영역을 제외하고는 C, D 영역에 해당하므로 신뢰도가 매우 높다고 판단할 수 있다. 15/9-F-1A의 DT를 예측한 다른 모델 결과인 Figure 3.20과 Figure 3.24에서 A, B 영역인 부분이 Figure 3.28 에서는 신뢰도가 높은 C, D 영역으로 변화된 것을 확인할 수 있었다.

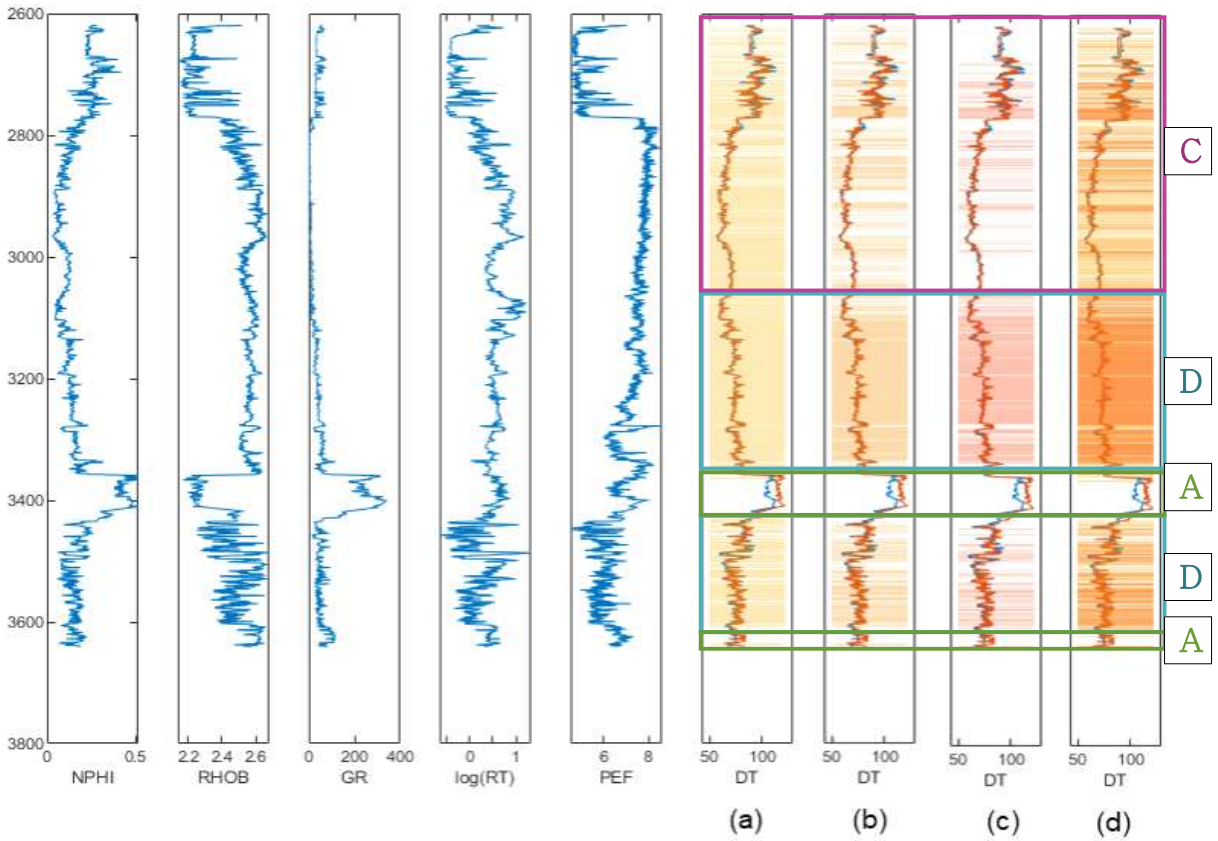


Figure 3.28 Well log of input data and predicted DT of RF_1B+11A_1A_model for g value change.

2. 15/9-F-11A 예측 모델 결과

1) 15/9-F-1B를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.29(a)와 (b)는 15/9-F-1B 데이터를 기준으로 SVDD를 사용해 g=1로 설정하였을 때 생성되는 경계를 기준으로 15/9-F-11A의 DT를 예측한 결과이다. outBND 모델의 예측 결과를 보면 실제값과 예측값이 잘 매칭되지 않아 보였다. 2차원 맵핑을 통해 본 Figure 3.29(d)의 outBND 데이터의 거리 또한 멀리 떨어져 분포했다.

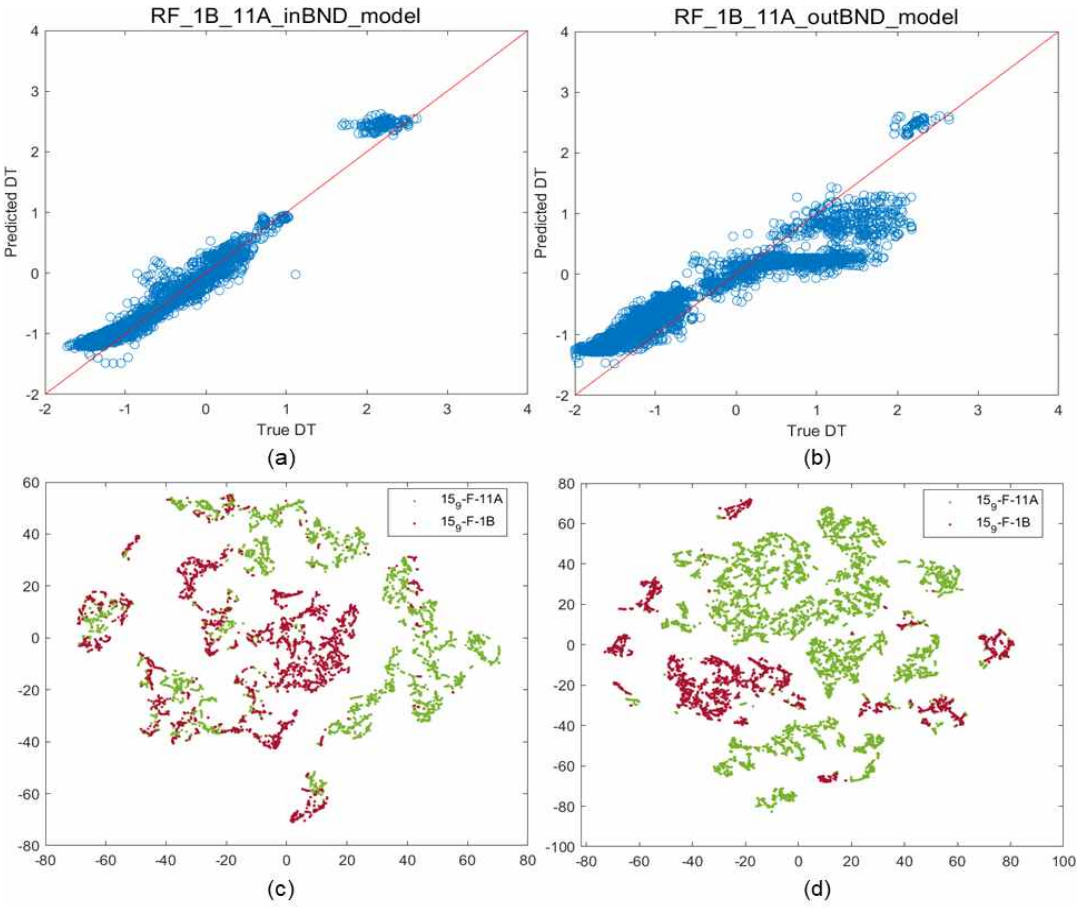


Figure 3.29 Prediction result of RF_1B_11A_inBND and RF_1B_11A_outBND model for g=1.

Figure 3.30(b)와 같이 3D 형태로 보면 경계에 데이터가 분포되어 있는 것을 확인할 수 있다. Figure 3.30(a)는 데이터를 분리하기 전의 15/9-F-1B와 15/9-F-11A 분포형태이며, Figure 3.30(c)는 outBND 데이터 분포로 inBND과 확실한 차이를 보인다. inBND 데이터는 학습데이터의 입력자료와 예측 데이터의 입력자료가 거의 겹쳐있어 예측 성능이 좋을 것으로 예상되며, outBND 데이터는 거의 분리되어있어 예측 오차가 inBND 데이터보다 클 것으로 예상되었다.

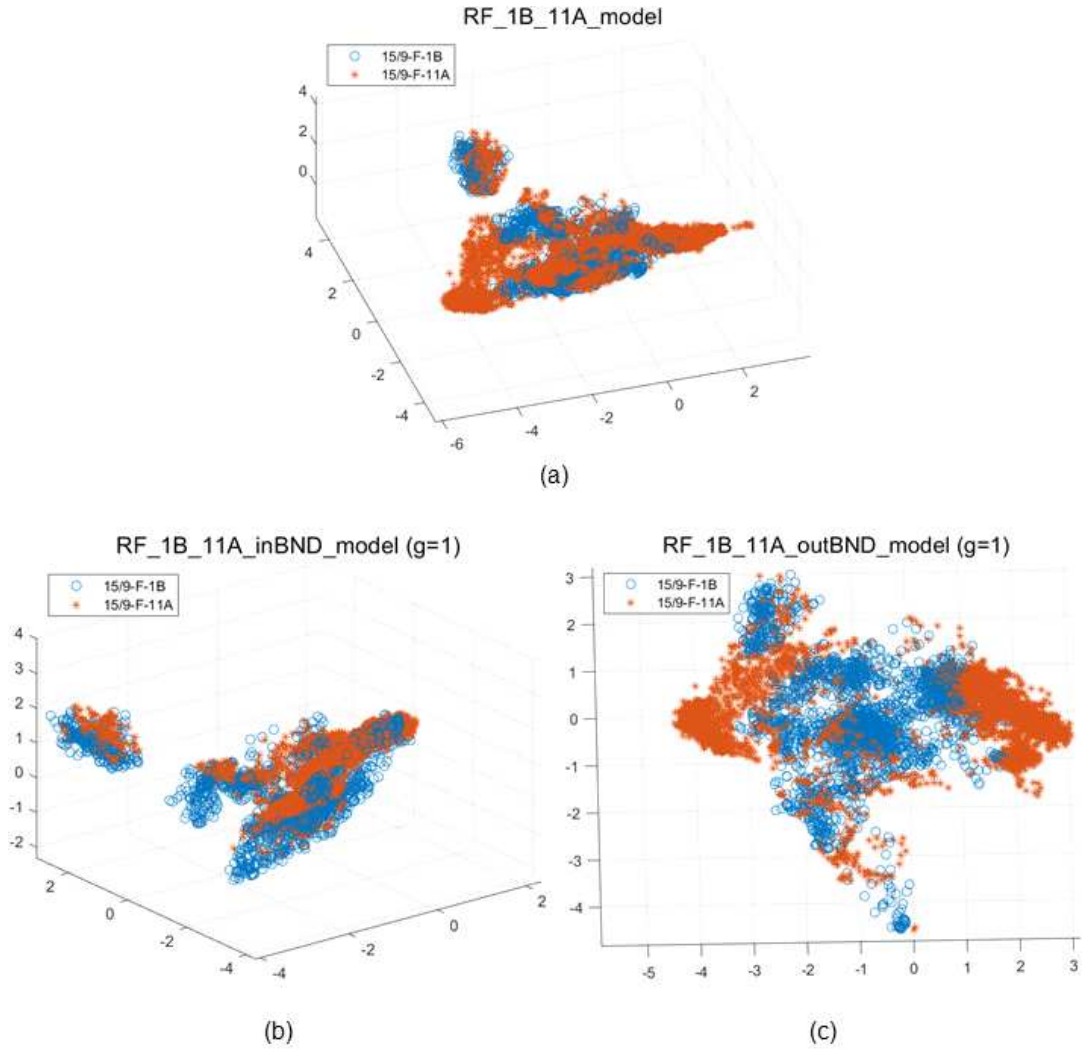


Figure 3.30 Result of PCA of RF_1B_11A_model, inBND_model and outBND_model for g=1.

나. g=3인 경우

Figure 3.31(a)와 (b)는 15/9-F-1B 데이터를 기준으로 SVDD를 사용해 g=3으로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-11A의 DT를 예측한 결과이다. Figure 3.31(a)는 g=1일 때의 결과보다 예측이 잘 된 것으로 보인다. inBND 영역을 나타낸 Figure 3.31(c)는 Figure 3.29(c)와는 다르게 15/9-F-1B와 거리가 가까운 15/9-F-11A 데이터만 남아 있는 것을 확인할 수 있다.

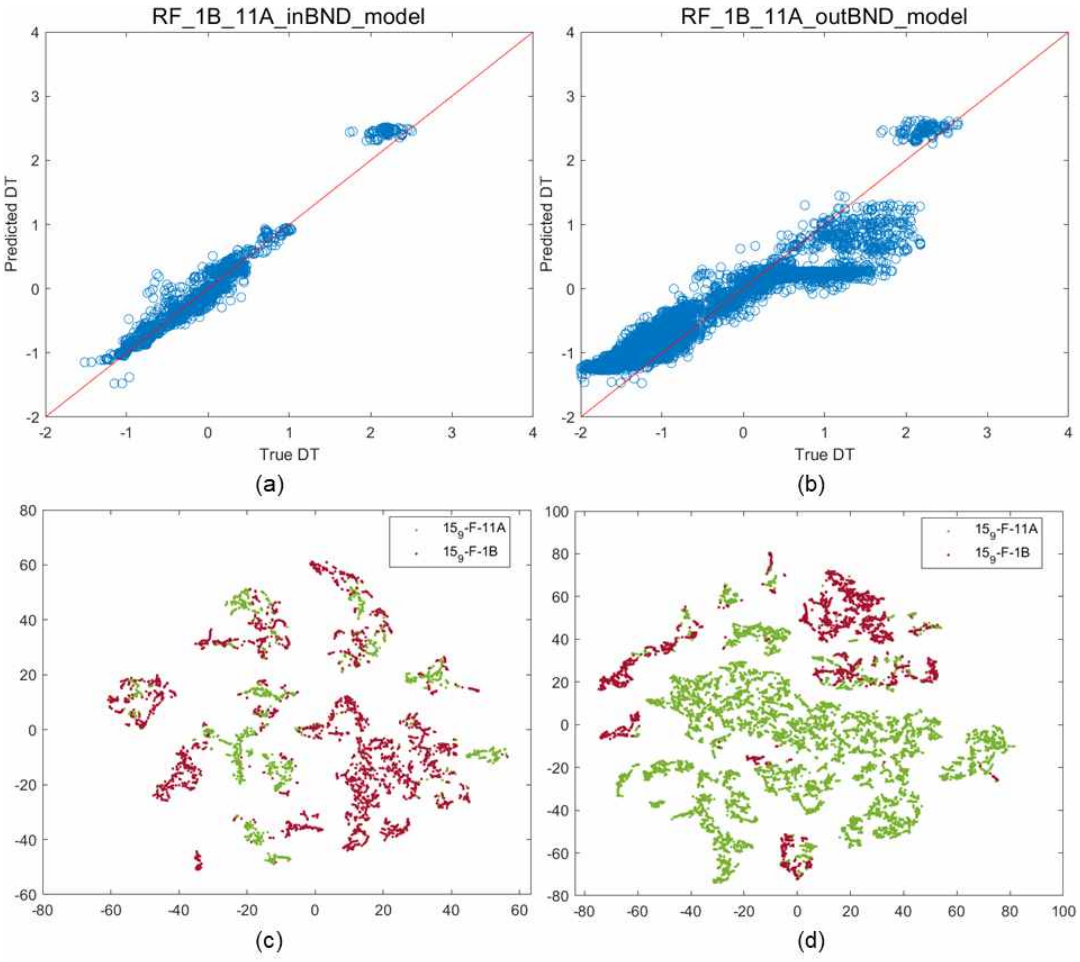


Figure 3.31 Prediction result of RF_1B_11A_inBND and RF_1B_11A_outBND model for g=3.

다. g=5인 경우

Figure 3.32는 g=5로 설정하였을 때 15/9-F-11A의 DT를 예측한 결과이다. Figure 3.32(a) 그래프에서 확인할 수 있듯 아주 적은 양의 데이터만인 inBND영역에 남아있고 대부분의 데이터가 Figure 3.32(b) 그래프와 같이 outBND에 존재하는 것을 확인할 수 있다. inBND영역을 나타내는 Figure 3.32(c)는 Figure 3.31(c)와 비교하여 아주 적은 양인 것을 볼 수 있고 대부분의 데이터가 outBND 영역에 존재하는 것을 확인할 수 있다.

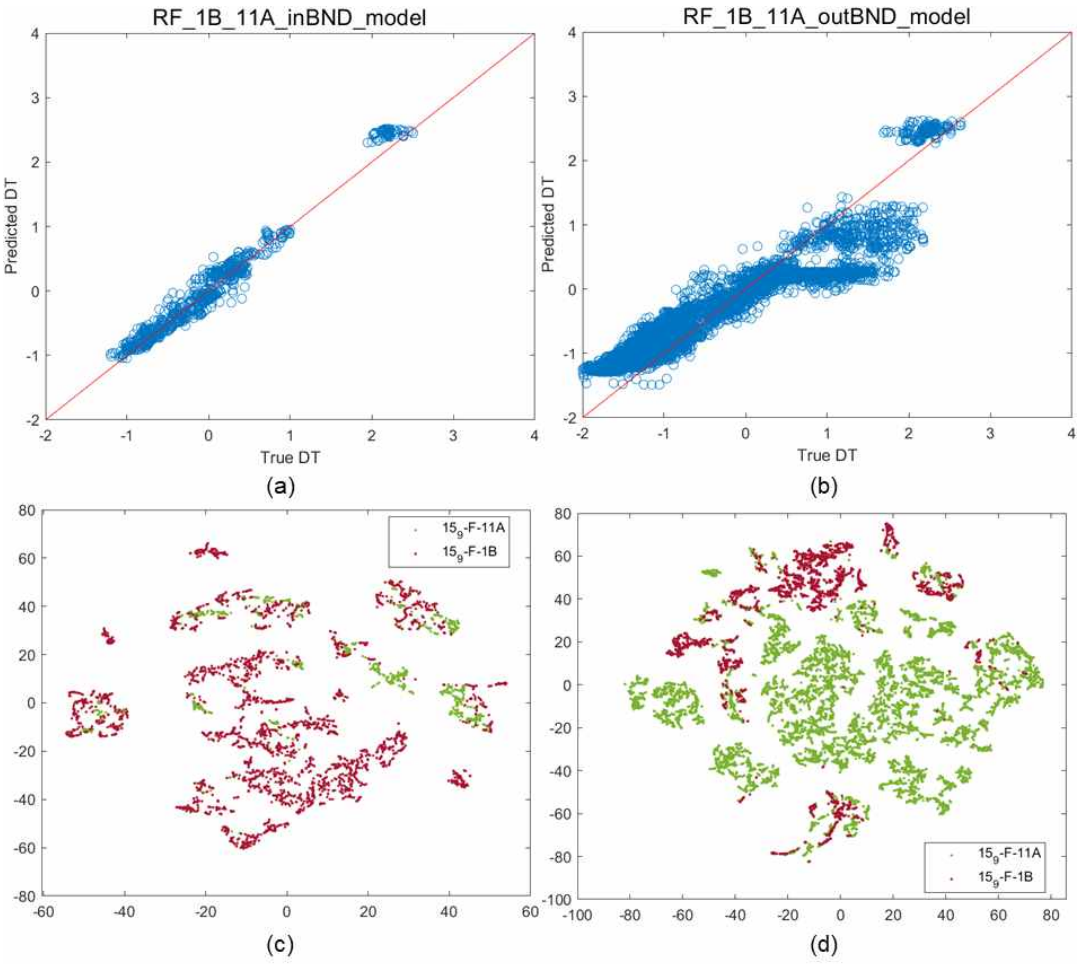


Figure 3.32 Prediction result of RF_1B_11A_inBND and RF_1B_11A_outBND model for g=5.

Figure 3.33는 15/9-F-1B 데이터를 학습하여 15/9-F-11A의 DT를 예측한 모델의 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.4186으로 타 모델에 비해 높은 편이다.

inBND RMSE는 g=1일 때 0.1802, g=3일 때 0.1496, g=5일 때 0.1402로 점점 오차가 감소하는 것으로 나타났으며, 전체 오차보다 현저히 낮은 오차를 보여 데이터를 잘 분리했음을 알 수 있다. outBND RMSE는 g=1일 때 0.493, g=3일 때 0.4478, g=5일 때 0.4339로 전체 데이터를 가지고 학습할 때보다 조금 높은 오차를 보였다.

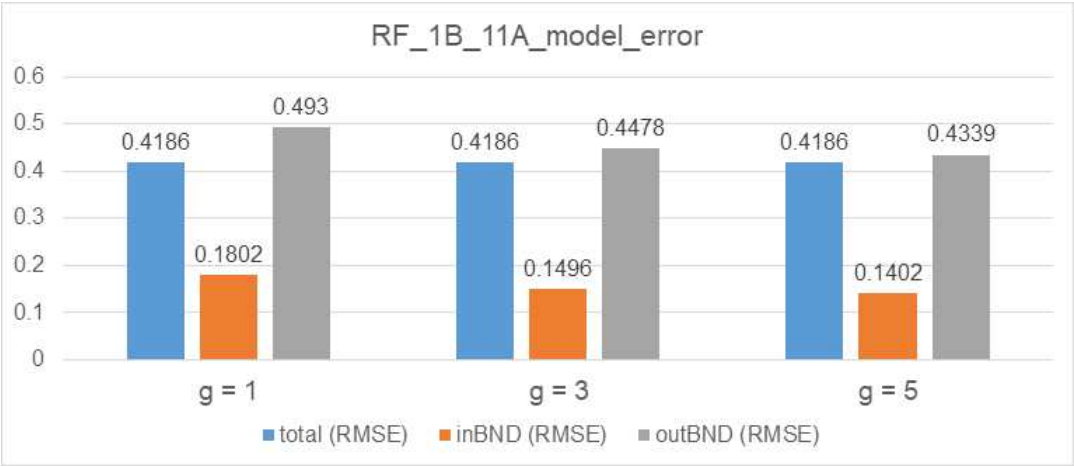


Figure 3.33 Comparison result of RF_1B_11A_model for g value change.

Figure 3.34는 오차분포의 빈도수를 나타낸 히스토그램이다. g=1, 5일 때 inBND 데이터 오차를 나타낸 Figure 3.34(a), (c)는 0을 중심으로 오차가 분포하여 예측값의 분포가 나쁘지 않음을 알 수 있다. 그러나 outBND 데이터 오차를 나타낸 Figure 3.34(b), (d)는 실제값보다 큰 값을 예측한 결과가 많음을 알 수 있으며, 이것은 Figure 3.29(b)와 Figure 3.32(b)에서도 확인할 수 있다.

Figure 3.34은 15/9-F-1B 데이터를 학습하여 15/9-F-11A를 예측하는 모델을 g=1, g=5인 경우 inBND와 outBND 데이터의 오차의 빈도수를 히스토그램으로 나타낸 그래프이다. inBND 데이터에 오차가 0에 가까운 빈도수가 outBND 데이터보다 많으며, g=5인 경우 inBND 데이터에 있던 0에 가까운 오차가 outBND 데이터에 포함된 것을 확인하였다.

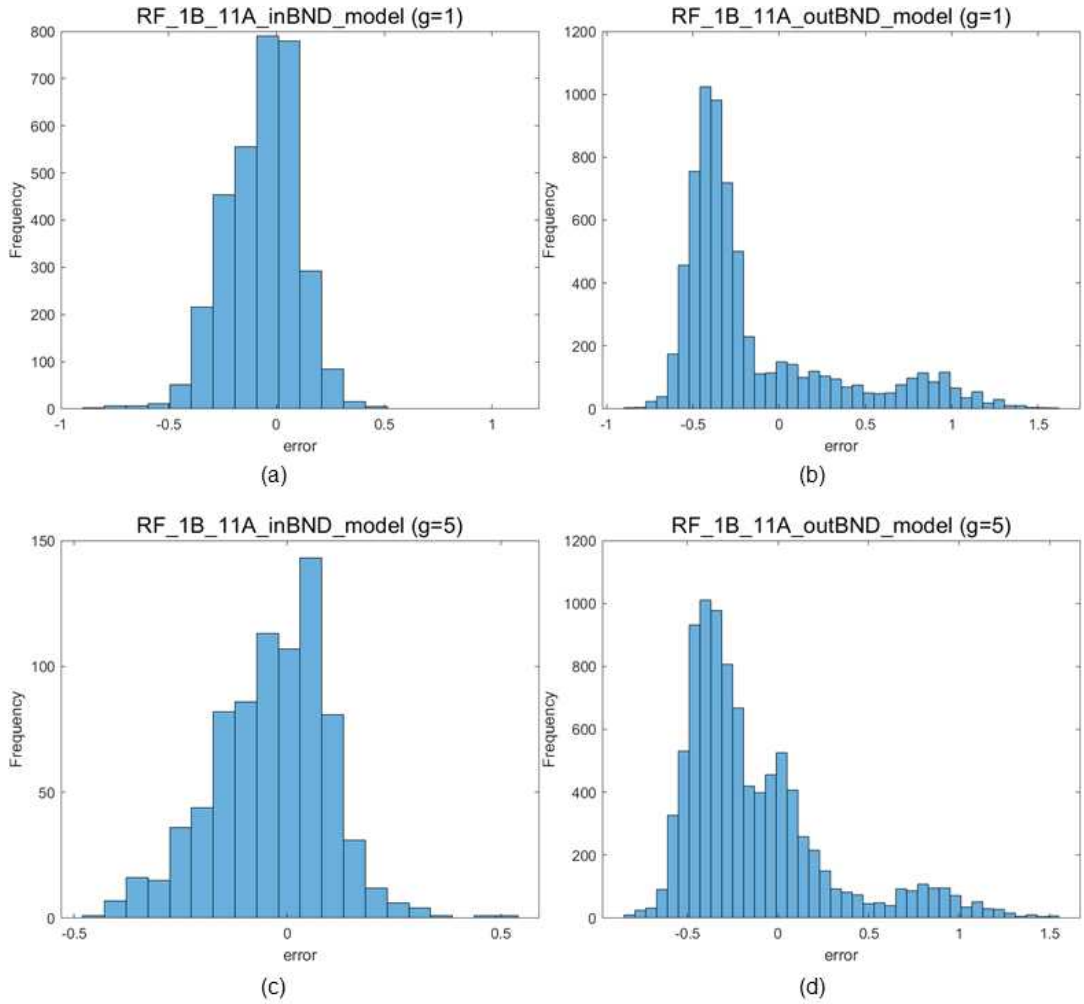


Figure 3.34 Histogram of RF_1B_11A_inBND_model error and RF_1B_11A_outBND_model error for g value change.

Table 3.8은 15/9-F-1B 데이터를 학습하여 15/9-F-11A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각각 32.23%, 14.21%, 7.75%로 점점 감소하였으며, 경계 외부에 해당하는 데이터의 비율은 각각 67.77%, 85.79%, 92.25%로 증가하였다. g값을 증가시키면 7.75%에서 32.23%만 inBND에 포함됨으로써 두 개의 시추공의 데이터 대부분은 유사성이 낮은 것으로 짐작할 수 있다.

전체 RMSE 대비 inBND RMSE는 상대적으로 g=1, 3, 5 일 때 56.95%, 64.26%, 66.51%로 감소하였으며, outBND RMSE는 g=1, 3, 5 일 때 17.77%, 6.98%, 3.66%로 소폭 증가하였다.

Table 3.8 Comparison result of RF_1B_11A_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	32.23 %	14.21 %	7.75 %
ratio of data out of boundary	67.77 %	85.79 %	92.25 %
(1)total RMSE	0.4186		
(2)inBND RMSE	0.1802	0.1496	0.1402
(3)outBND RMSE	0.493	0.4478	0.4339
relative difference between (1) and (2)	-56.95 %	-64.26 %	-66.51 %
relative difference between (1) and (3)	17.77 %	6.98 %	3.66 %

Figure 3.35는 15/9-F-11A 데이터의 NPHI, RHOB, GR, RT, PEF와 예측된 DT를 Depth에 따라 그래프로 나타낸 것이다. 실제 DT는 파란색 선이고 예측된 DT는 주황색 선으로 나타냈다. Figure 3.35(a)의 색칠된 영역은 g=1인 경우로 전체 데이터의 32.23%에 해당하는 inBND 데이터의 예측 결과에 해당한다. Figure 3.35(b)의 색칠된 영역은 g=3인 경우로 전체 데이터의 14.21 %에 해당하는 inBND 데이터

의 예측 결과이다. 해당 영역의 오차는 전체 오차보다 64.26% 낮다. Figure 3.35(c)의 색칠된 영역은 g=5인 경우로 전체 데이터의 7.75%이며, 해당 영역의 오차는 전체 오차보다 66.51% 낮다. Figure 3.35(d)는 (a), (b)와 (c)를 모두 함께 나타낸 결과이다.

g=1, 3, 5로 변화시켜가면서 경계영역 내에 존재하는 데이터를 분석하면 예측결과에 대해 높은 신뢰도를 보이는 구간을 선정할 수 있다. Figure 3.35는 예측 구간의 절반 이상이 A, B 영역으로 분포하였고, 나머지는 C 영역으로 나타났다.

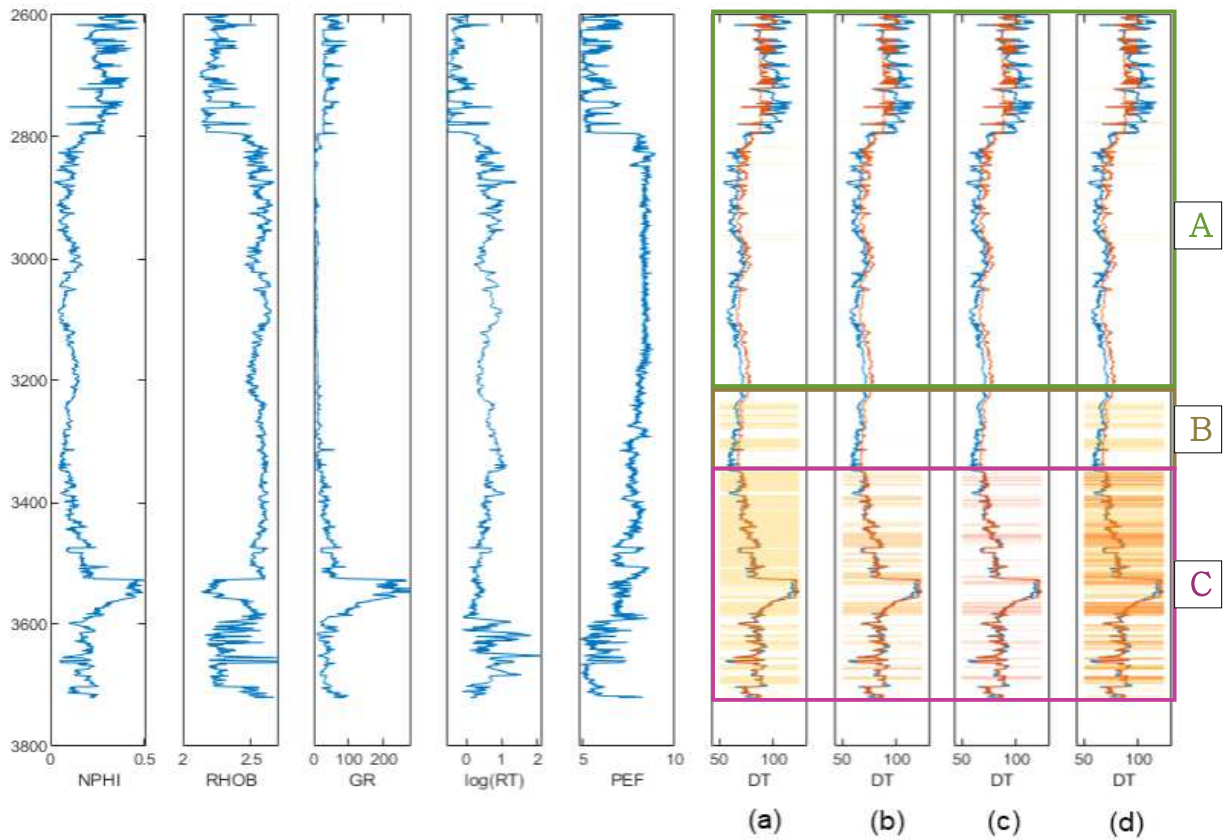


Figure 3.35 Well log of input data and predicted DT of RF_1B_11A_model for g value change

2) 15/9-F-1A를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.36(a)와 (b)는 15/9-F-1A 데이터를 기준으로 SVDD를 사용해 g=1으로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-11A의 DT를 예측한 결과이다. 실제값보다 작게 예측된 값이 outBND에 포함된 것을 확인할 수 있다.

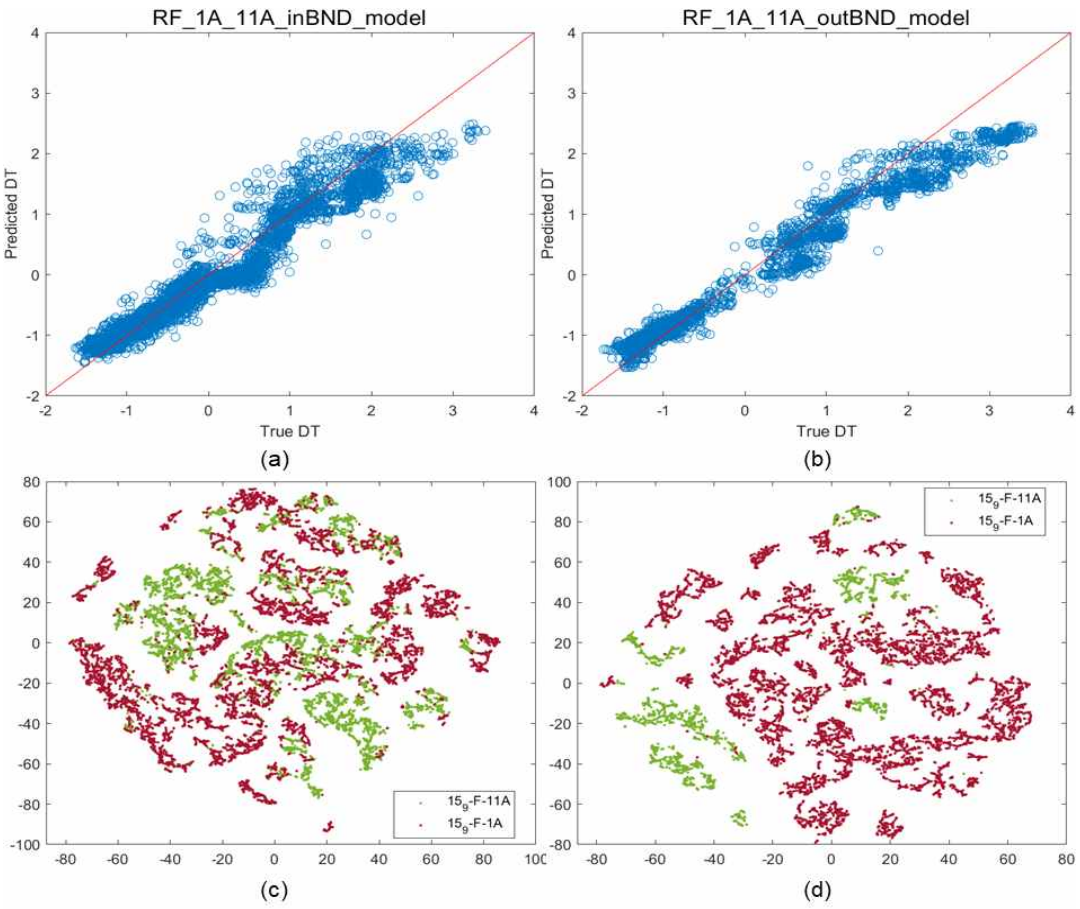


Figure 3.36 Prediction result of 15/9-F-11A using 15/9-F-1A training model for g=1.

나. g=5인 경우

Figure 3.36(a)와 (b)는 15/9-F-1A 데이터를 기준으로 SVDD를 사용해 g=5으로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-11A의 DT를 예측한 결과이다. g=1인 경우의 결과인 Figure 3.36(a), (b)와 해당 결과를 비교해봤을 때, 실제값보다 작게 예측된 값이 추가적으로 outBND에 포함된 것을 확인할 수 있었다.

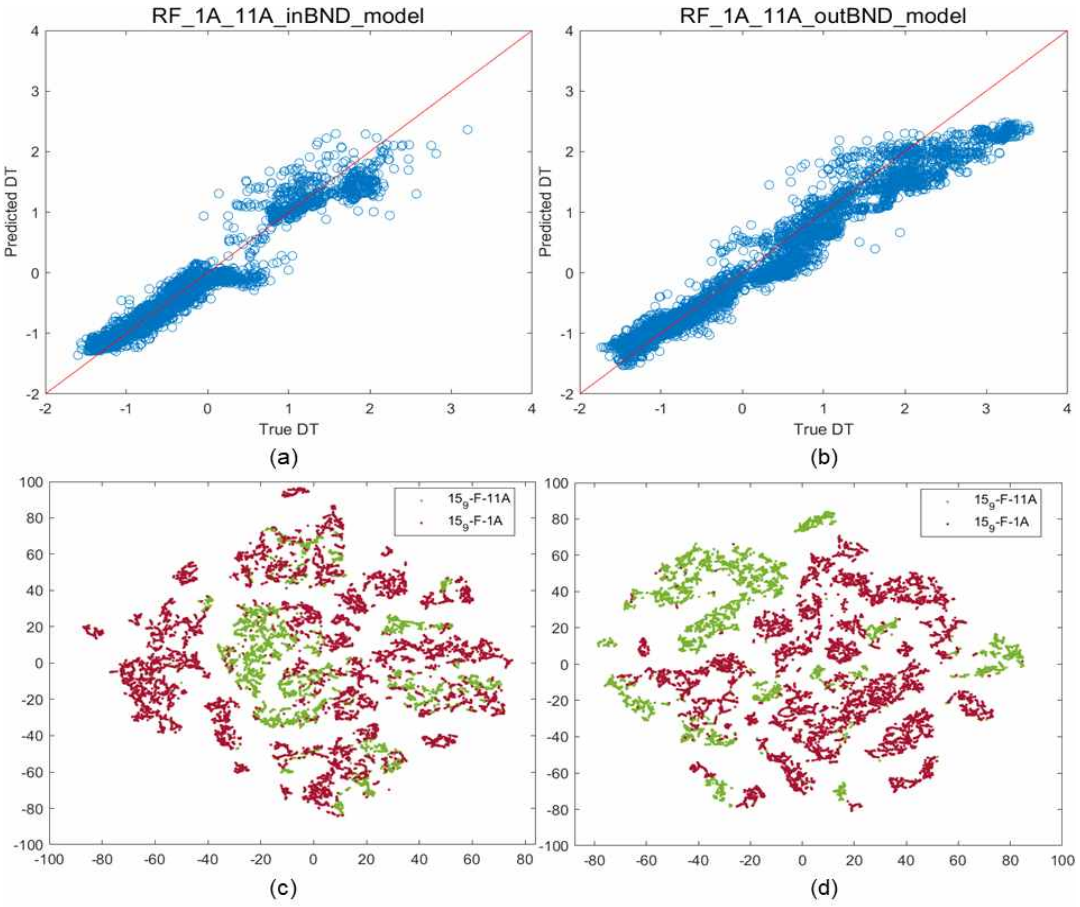


Figure 3.37 Prediction result of 15/9-F-11A using 15/9-F-1A training model for g=5.

Figure 3.38은 15/9-F-1A 데이터를 학습하여 15/9-F-11A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.2801이다.

inBND RMSE는 g=1일 때 0.2336, g=3일 때 0.2062, g=5일 때 0.1974로 점점 오차가 감소하는 것으로 나타났다. 이는 전체 데이터를 가지고 학습한 모델보다 적은 오차로 나타났으며, g=5일 때의 오차가 가장 적었다. 반면, outBND RMSE는 g=1일 때 0.3868, g=3일 때 0.3563, g=5일 때 0.3382로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다.

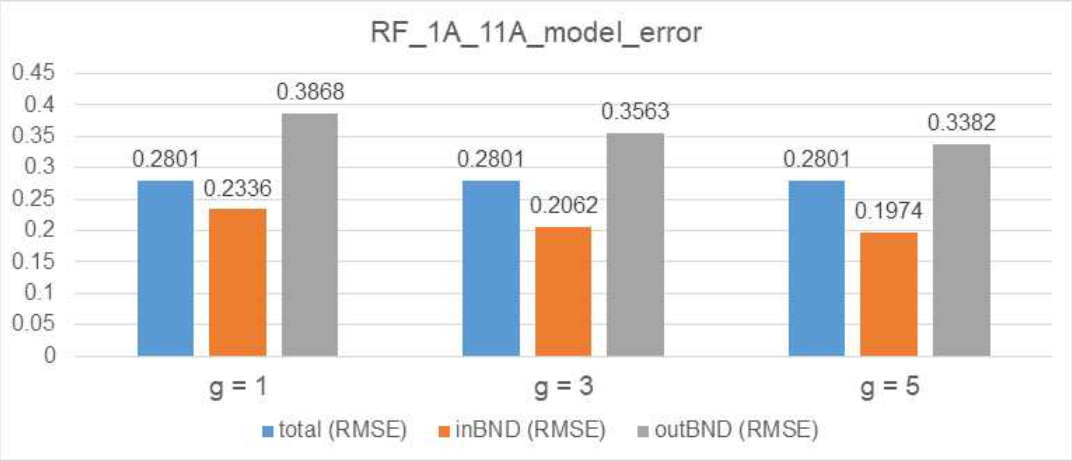


Figure 3.38 Comparison result of RF_1A_11A_model for g value change.

Table 3.9는 15/9-F-1A 데이터를 학습하여 15/9-F-11A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각 각 74.85%, 57.39%, 47.62%이며, 경계 외부에 해당하는 데이터의 비율은 25.15%, 42.61%, 52.38%로 증가하였다.

전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 16.60%, g=3일 때, 26.38%, g=5일 때, 29.53% 감소하였으며, outBND RMSE는 g=1일 때, 38.09%, g=3일 때, 27.20%, g=5일 때, 20.74% 증가하였다.

Table 3.9 Relative difference result of RF_1A_11A_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	74.85 %	57.39 %	47.62 %
ratio of data out of boundary	25.15 %	42.61 %	52.38 %
(1)total RMSE	0.2801		
(2)inBND RMSE	0.2336	0.2062	0.1974
(3)outBND RMSE	0.3868	0.3563	0.3382
relative difference between (1) and (2)	-16.60 %	-26.38 %	-29.53 %
relative difference between (1) and (3)	38.09 %	27.20 %	20.74 %

Figure 3.39는 15/9-F-1A 데이터를 학습한 모델로 15/9-F-11A의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간 중 일부분이 A, B 영역에 해당하지만, 해당 영역을 제외하고는 C, D 영역에 해당하므로 신뢰도가 매우 높다고 판단할 수 있다. 15/9-F-1B 데이터를 학습하여 예측한 결과인 Figure 3.35에서 A, B 영역이 해당 결과에서 C, D 영역으로 변화된 것을 확인할 수 있었다. 또한 Figure 3.35의 C 영역의 일부가 해당 결과에서는 A, B 영역으로 나타났다.

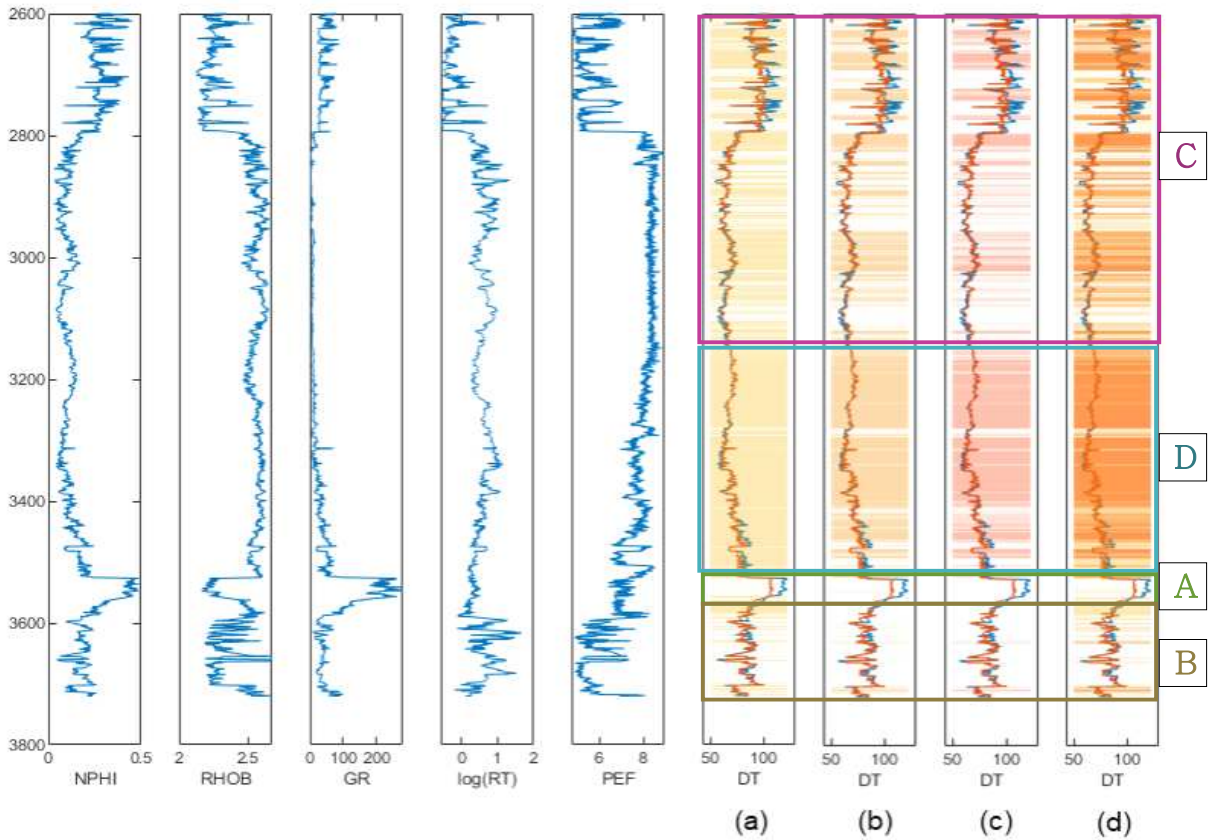


Figure 3.39 Well log of input data and predicted DT of RF_1A_11A_model for g value change

3) 15/9-F-1A와 15/9-F-1B를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.40(a)와 (b)는 15/9-F-1A와 15/9-F-1B 데이터를 기준으로 SVDD를 사용해 g=1으로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-11A의 DT를 예측한 결과이다. Figure 3.40(c), (d)를 보면 inBND 데이터가 outBND 데이터보다 거리가 가까움을 확인하였다.

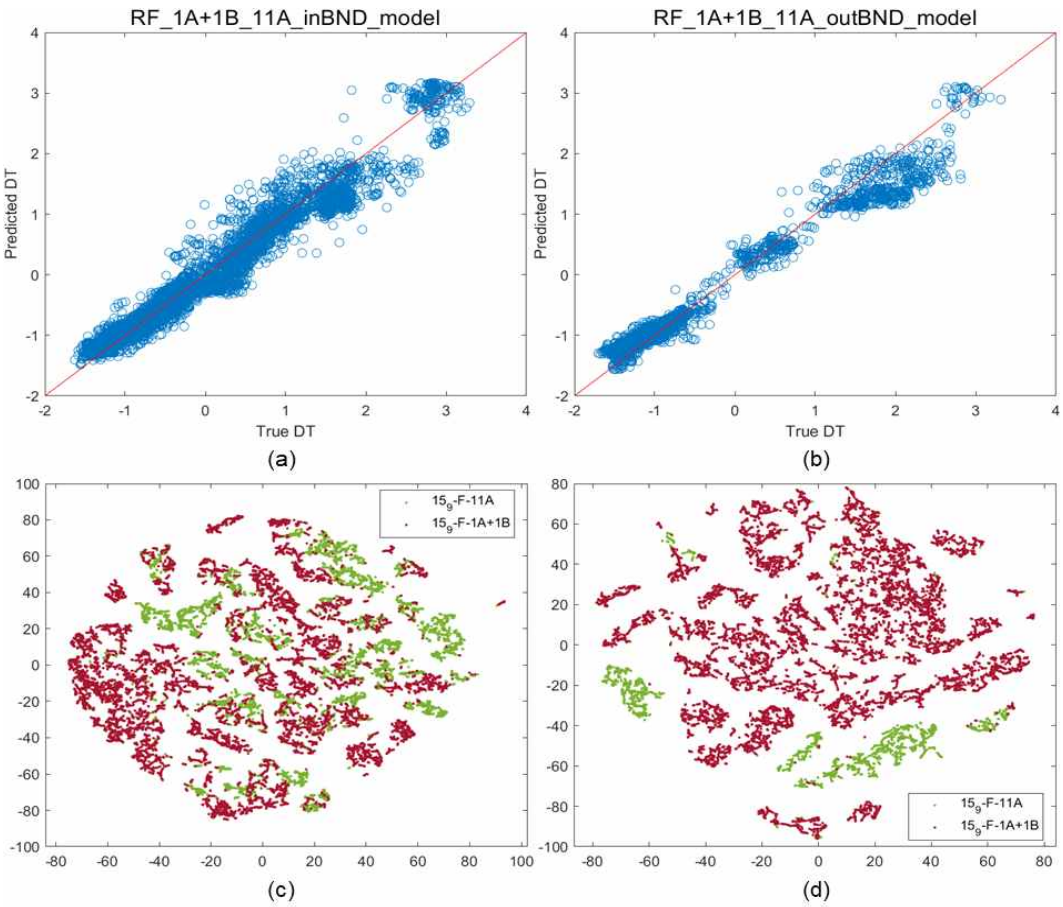


Figure 3.40 Prediction result of 15/9-F-11A using 15/9-F-1A and 15/9-F-1B training model for g=1.

나. g=5인 경우

Figure 3.41(a)와 (b)는 15/9-F-1A와 15/9-F-1B 데이터를 기준으로 SVDD를 사용해 g=5로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-11A의 DT를 예측한 결과이다. Figure 3.40(a), (b)와 비교해봤을 때, g를 5로 증가시키자 inBND에서 예측이 잘 맞지 않는 데이터가 outBND로 이동한 것을 확인할 수 있었다.

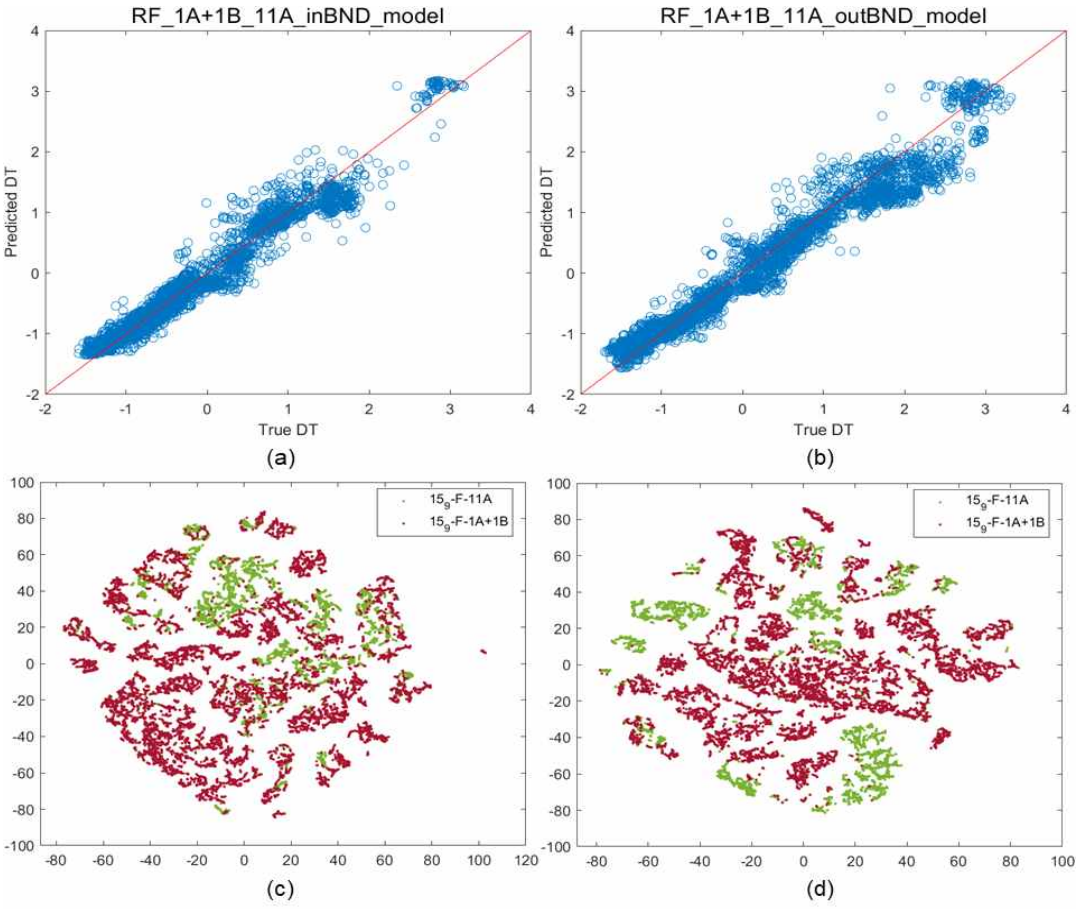


Figure 3.41 Prediction result of 15/9-F-11A using 15/9-F-1A and 15/9-F-1B training model for g=5.

Figure 3.42는 15/9-F-1A와 15/9-F-1A 데이터를 학습하여 15/9-F-11A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.2182이다.

inBND RMSE는 g=1일 때 0.2027, g=3일 때 0.1916, g=5일 때 0.1772로 점점 오차가 감소하는 것으로 나타났다. 이는 전체 데이터를 가지고 학습한 모델보다 작은 오차로 나타났으며, g=5일 때의 오차가 가장 작았다. 반면, outBND RMSE는 g=1일 때 0.2649, g=3일 때 0.2604, g=5일 때 0.2539로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다.

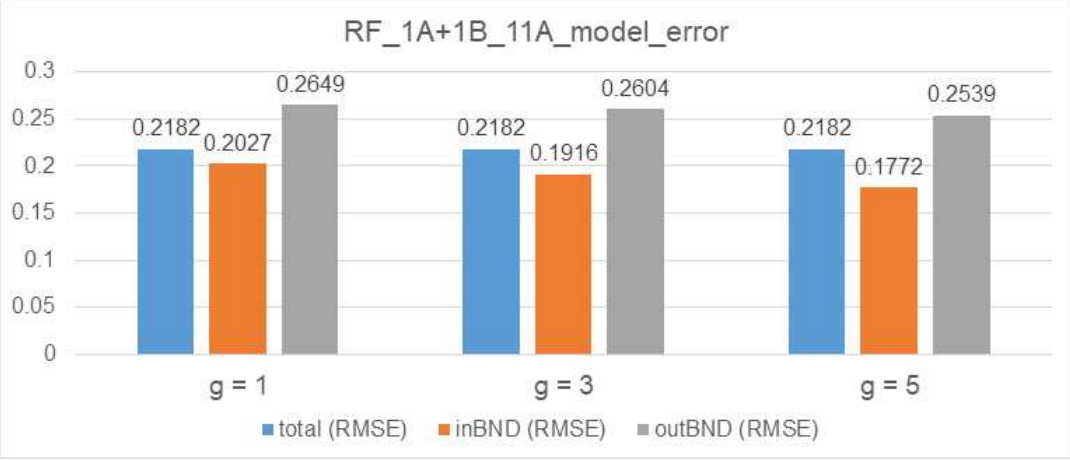


Figure 3.42 Comparison result of RF_1A+1B_11A_model for g value change.

Table 3.10은 15/9-F-1A와 15/9-F-1B 데이터를 학습하여 15/9-F-11A의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각각 77.51%, 63.80%, 50.93%이며, 경계 외부에 해당하는 데이터의 비율은 22.49%, 36.20%, 49.07%로 증가하였다.

전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 7.10%, g=3일 때, 12.19%, g=5일 때, 18.79% 감소하였으며, outBND RMSE는 g=1일 때, 21.40%, g=3일 때, 19.34%, g=5일 때, 16.36% 증가하였다.

Table 3.10 Relative difference result of RF_1A+1B_11A_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	77.51 %	63.80 %	50.93 %
ratio of data out of boundary	22.49 %	36.20 %	49.07 %
(1)total RMSE	0.2182		
(2)inBND RMSE	0.2027	0.1916	0.1772
(3)outBND RMSE	0.2649	0.2604	0.2539
relative difference between (1) and (2)	-7.10 %	-12.19 %	-18.79 %
relative difference between (1) and (3)	21.40 %	19.34 %	16.36 %

Figure 3.43은 15/9-F-1A와 15/9-F-1B 데이터를 학습한 모델로 15/9-F-11A의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간은 C, D 영역으로 분포하여 신뢰도가 매우 높다고 판단할 수 있다. 15/9-F-1B 데이터를 학습하여 예측한 결과인 Figure 3.35에서 A, B 영역과 15/9-F-1A 데이터를 학습하여 예측한 결과인 Figure 3.39에서 A, B 영역이 해당 결과에서 C, D 영역으로 변화된 것으로 나타났다. 따라서 하나의 시추공 데이터를 학습하여 예측한 결과보다 두 개의 시추공 데이터를 학습하여 예측한 결과의 신뢰도가 더 높음을 확인할 수 있었다.

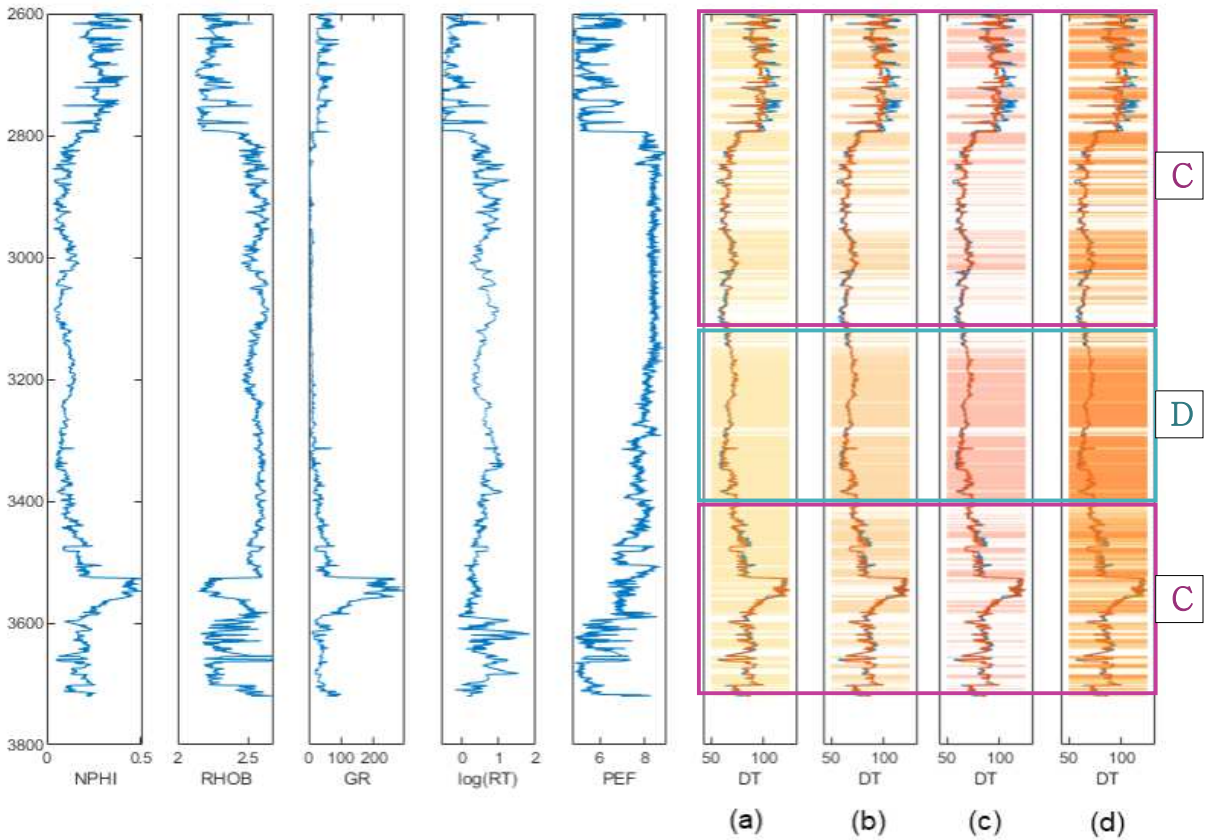


Figure 3.43 Well log of input data and predicted DT of RF_1A+1B_11A_model for g value change

3. 15/9-F-1B 예측 모델 결과

1) 15/9-F-1A와 15/9-F-11A를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.44(a)와 (b)는 15/9-F-1A와 15/9-F-11A 데이터를 함께 사용하여 SVDD를 사용해 g=1로 설정하였을 때 생성되는 경계를 기준으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-1B의 DT를 예측한 결과이다. 15/9-F-1B의 데이터의 개수가 2566개로 15/9-F-1A와 15/9-F-11A 데이터를 합한 개수보다 현저히 작고 학습데이터가 넓은 공간에 분포되어 대부분의 15/9-F-1B 데이터가 inBND영역에 속하는 것으로 나타났다. 대부분이 inBND영역에 분포해있기 때문에 outBND영역을 나타낸 Figure 3.44(d)에서는 데이터가 거의 보이지 않는다.

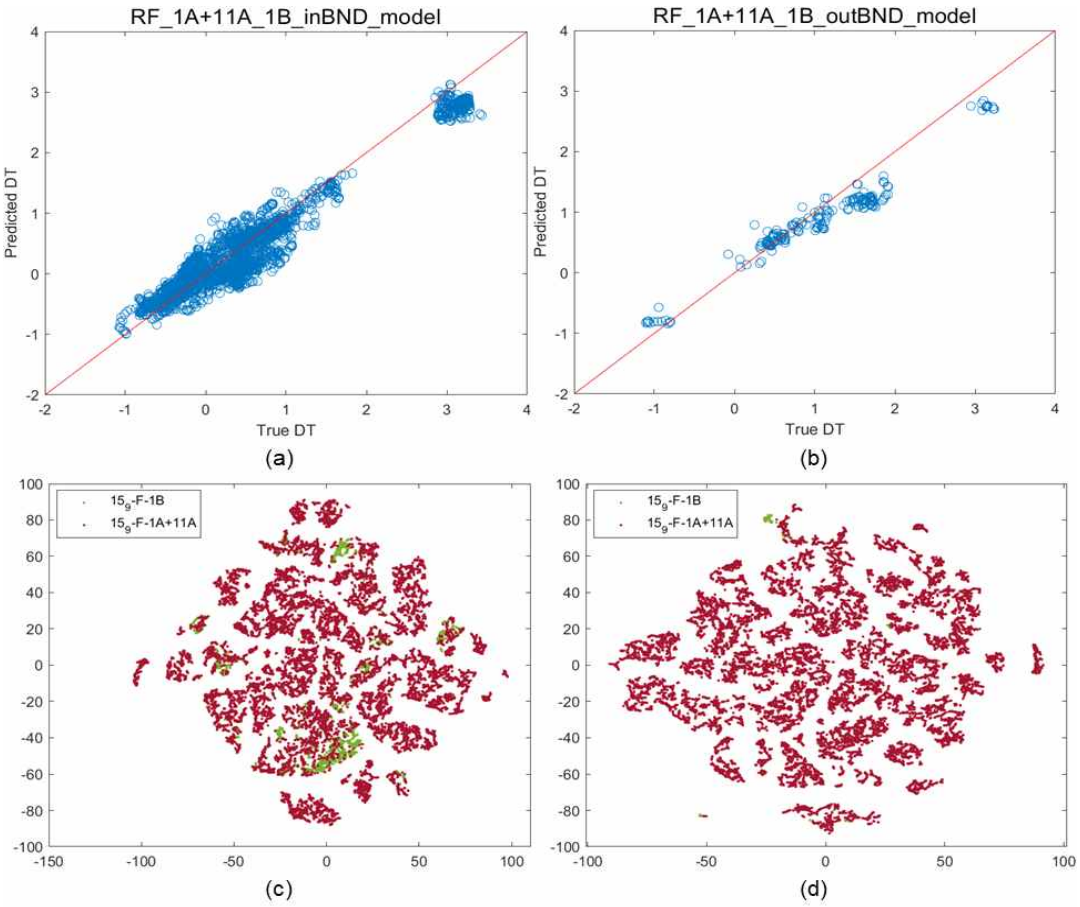


Figure 3.44 Prediction result of 15/9-F-1B using 15/9-F-1A and 15/9-F-11A training model for g=1.

Figure 3.45(b)와 같이 3D 형태로 보면 경계 내에 데이터가 분포되어 있는 것을 확인할 수 있다. Figure 3.45(a)는 데이터를 분리하기 전의 학습데이터인 15/9-F-1A, 15/9-F-1B의 입력자료와 예측 데이터 15/9-F-11A의 입력자료의 분포 형태이며, Figure 3.45(c)는 outBND 데이터 분포로 경계 외부에 데이터가 분포되어 있는 것을 확인할 수 있다.

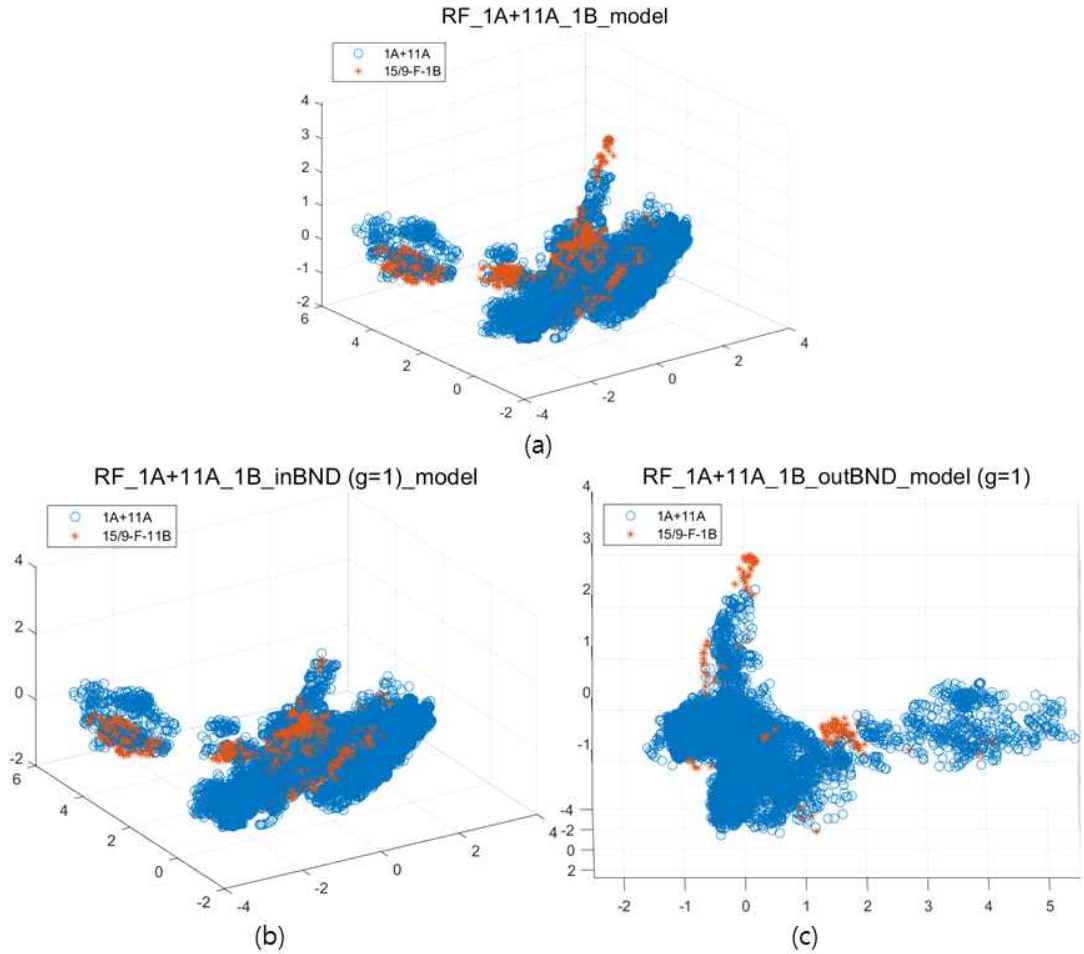


Figure 3.45 Result of PCA of RF_1B_11A_model, inBND_model and outBND_model for g=1.

나. g=3인 경우

Figure 3.46(a)와 (b)는 SVDD를 사용해 g=3로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-1B의 DT를 예측한 결과이다. Figure 3.44(b)에 비해 outBND 데이터 수가 증가한 것을 확인할 수 있다.

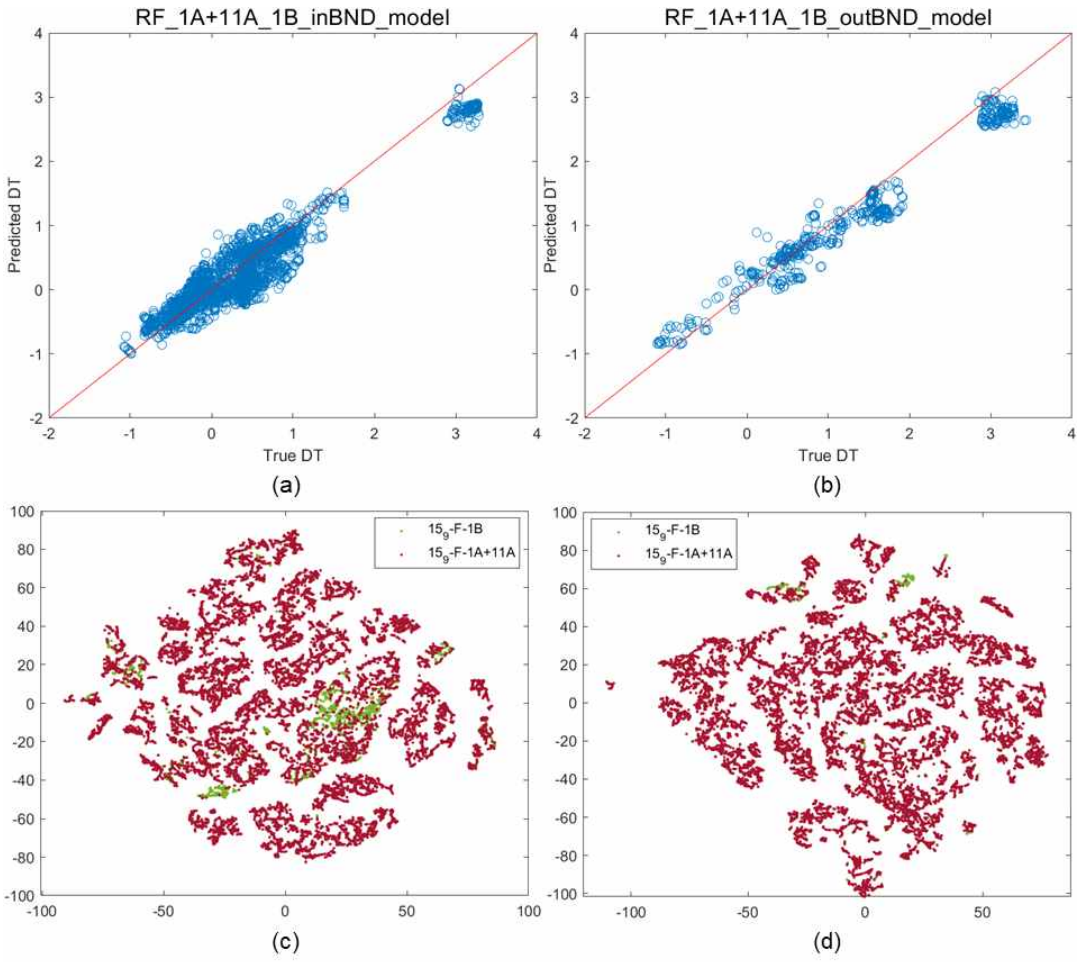


Figure 3.46 Prediction result of 15/9-F-1B using 15/9-F-1A and 15/9-F-11A training model for g=3.

다. g=5인 경우

Figure 3.47은 g=5로 설정하였을 때 15/9-F-1B의 DT를 예측한 결과이다. Figure 3.47(a)의 실제값 보다 작게 예측된 부분은 Figure 3.44(a), Figure 3.46(a)와 비교해보았을 때, Figure 3.47(b)의 outBND로 이동한 것을 확인할 수 있다. 학습 데이터의 개수는 19341개로 넓은 공간에 분포되어 있어 g값을 증가시켜도 15/9-F-1B의 69.68%가 inBND에 포함되었다.

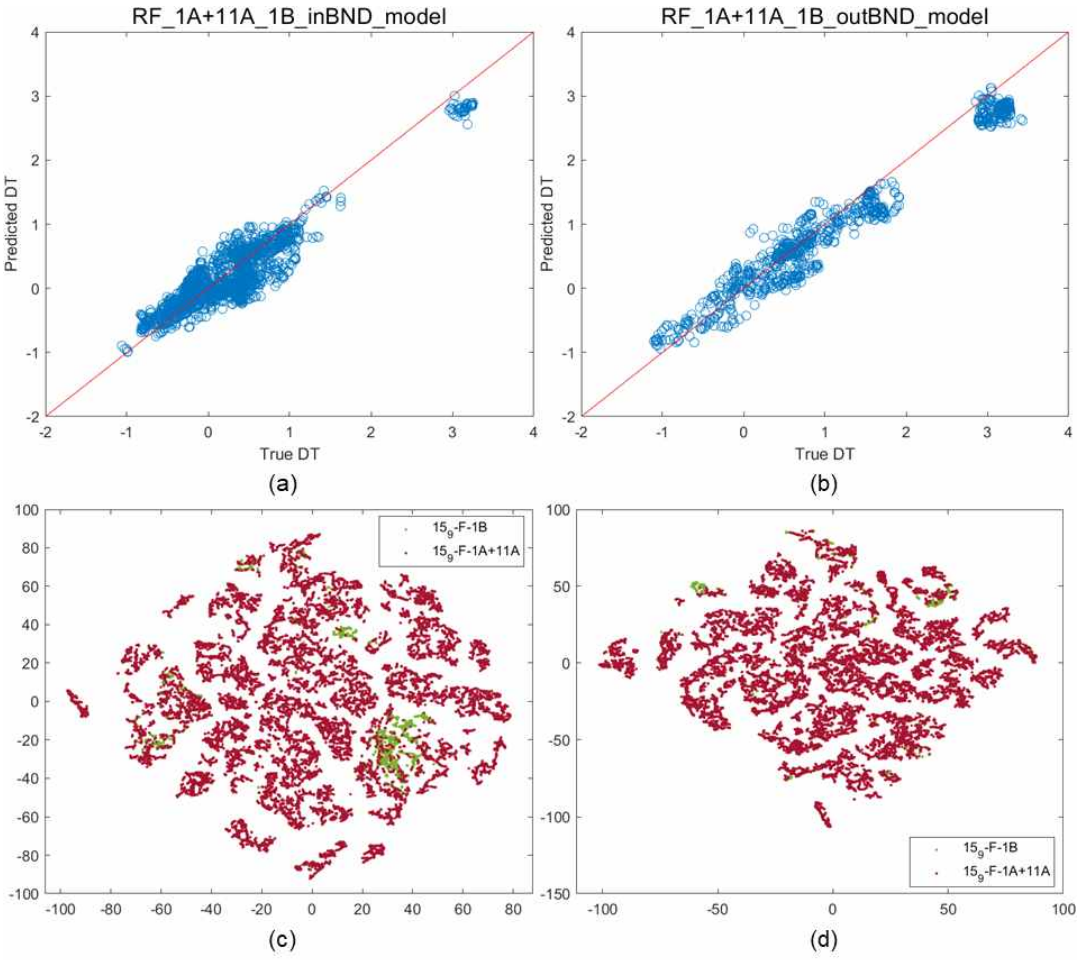


Figure 3.47 Prediction result of 15/9-F-1B using 15/9-F-1A and 15/9-F-11A training model for g=5.

Figure 3.48는 15/9-F-1A와 15/9-F-11A 데이터를 학습하여 15/9-F-1B의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.2663이다.

inBND RMSE는 g=1일 때 0.2607, g=3일 때 0.2551, g=5일 때 0.2497로 오차가 소폭 감소하는 것으로 나타났으며, outBND RMSE는 g=1일 때 0.3275, g=3일 때 0.3137, g=5일 때 0.3008로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다. g값이 증가함에 따라서 inBND와 outBND 오차가 모두 감소하는 경향을 보였다.

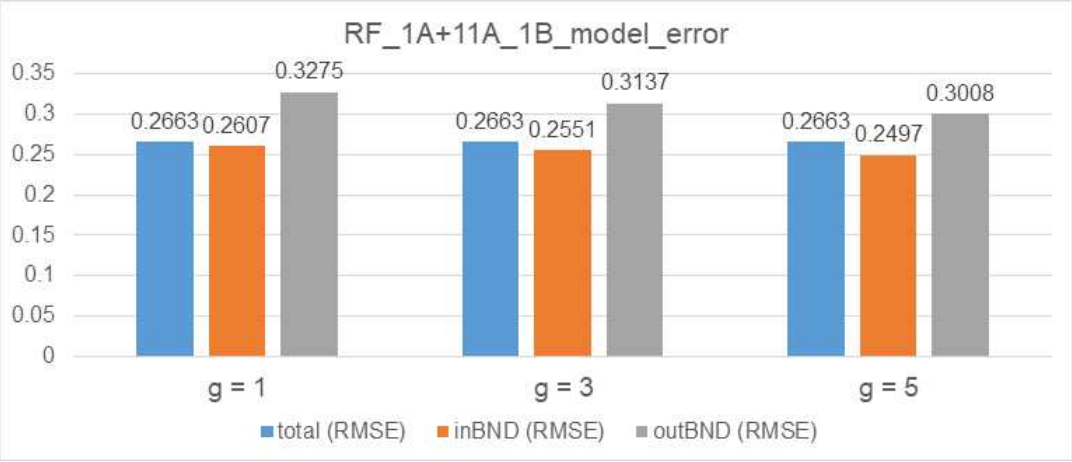


Figure 3.48 Comparison result of RF_1A+11A_1B_model for g value change.

Figure 3.49은 15/9-F-1A와 15/9-F-11A 데이터를 학습하여 15/9-F-1B를 예측하는 모델을 g=1, g=5인 경우 inBND와 outBND 데이터의 오차의 빈도수를 히스토그램으로 나타낸 그래프이다. inBND 데이터에 오차가 0에 가까운 빈도수가 outBND 데이터보다 많으며, g=5인 경우 inBND 데이터에 있던 0에 가까운 오차가 outBND 데이터에 포함된 것을 확인하였다.

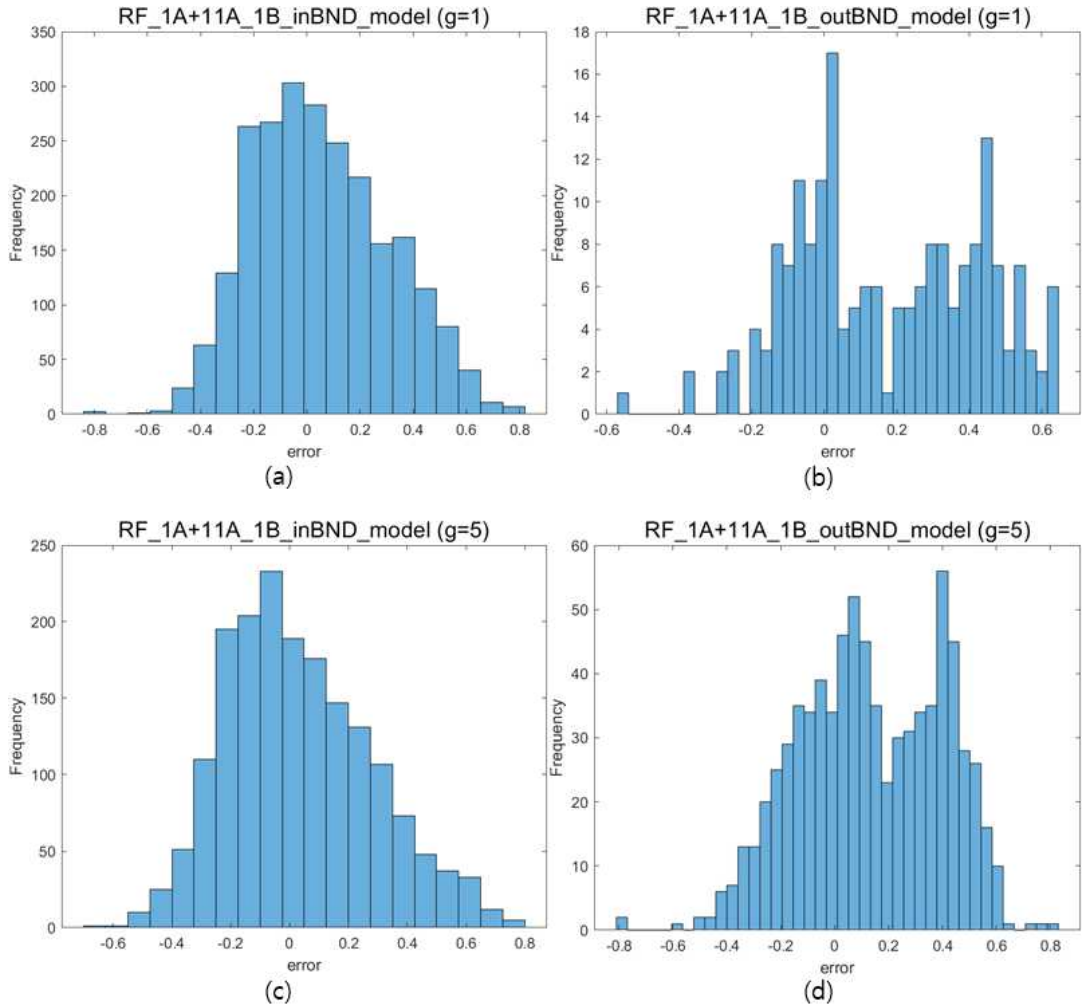


Figure 3.49 histogram of RF_1A+11A_1B_inBND_model error and RF_1A+11A_1B_outBND_model error for g value change.

Table 3.11은 15/9-F-1A와 15/9-F-11A 데이터를 학습하여 15/9-F-1B의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각각 92.52%, 82.54%, 69.68%로 대부분이 포함되었고, 경계 외부에 해당하는 데이터의 비율은 각각 7.48%, 17.46%, 30.32%로 증가하였다. g=5일 때, 절반이 넘는 데이터가 inBND에 포함되며, g=1일 때는 92.52%로 거의 대부분의 데이터가 포함되는 것을 확인할 수 있다. 이를 통해 두 데이터의 유사성이 높을 것으로 추측할 수 있으며, 2개의 시추공 데이터를 결합해서 학습데이터에 사용함으로써 더 높은 유사성을 나타내는 것으로 분석된다. 전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 2.10%, g=3일 때, 4.21%, g=5일 때, 6.23% 감소하였으며, outBND RMSE는 g=1일 때, 22.98%, g=3일 때, 17.80%, g=5일 때, 12.96% 증가하였다.

Table 3.11 Relative difference result of RF_1A+11A_1B_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	92.52 %	82.54 %	69.68 %
ratio of data out of boundary	7.48 %	17.46 %	30.32 %
(1)total RMSE	0.2663		
(2)inBND RMSE	0.2607	0.2551	0.2497
(3)outBND RMSE	0.3275	0.3137	0.3008
relative difference between (1) and (2)	-2.10 %	-4.21 %	-6.23 %
relative difference between (1) and (3)	22.98 %	17.80 %	12.96 %

Figure 3.50은 15/9-F-1A와 15/9-F-1B 데이터를 학습한 모델로 15/9-F-11A의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간은 일부분의 A 영역을 제외하고는 C, D 영역으로 분포하여 신뢰도가 매우 높다고 판단할 수 있다.

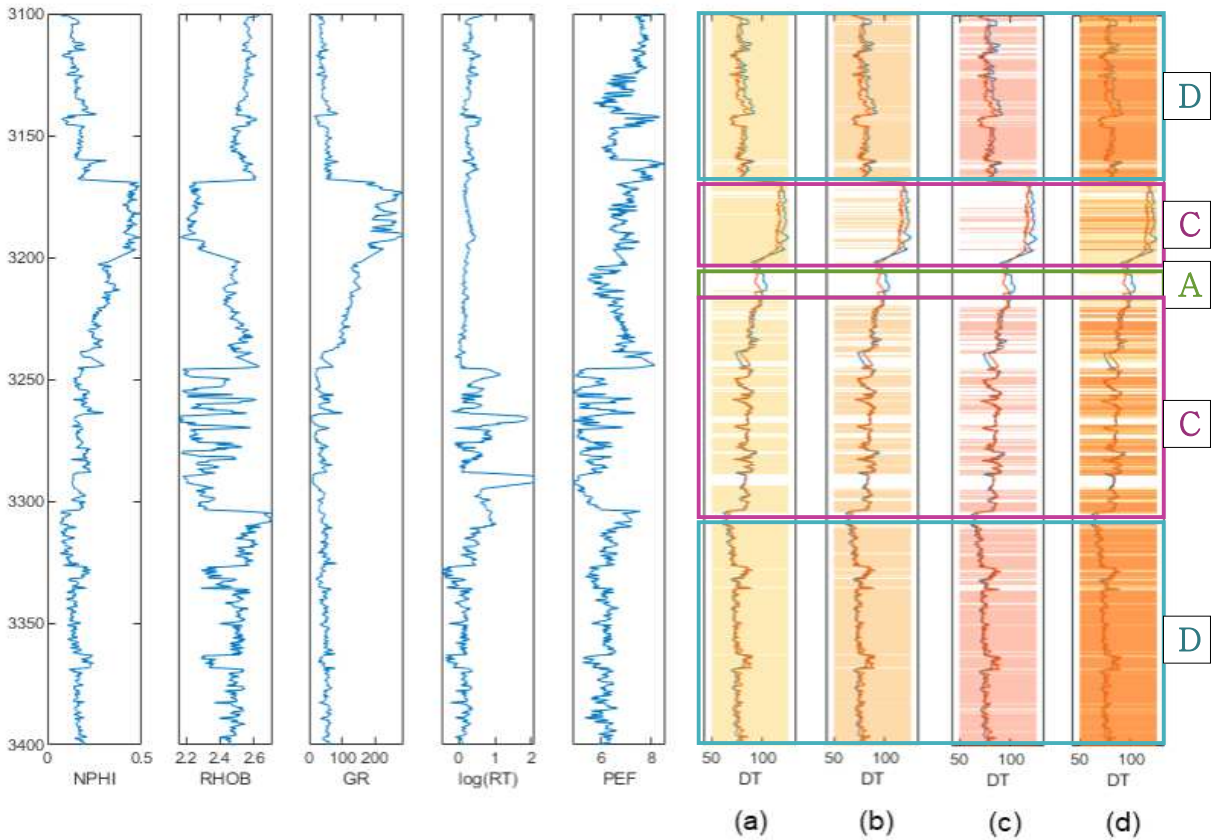


Figure 3.50 Well log of input data and predicted DT of RF_1A+11A_1B_model for g value change

2) 15/9-F-1A를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.51(a)와 (b)는 15/9-F-1A 데이터를 기준으로 SVDD를 사용해 g=1으로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-1B의 DT를 예측한 결과이다. 실제값보다 예측값이 작은 데이터가 outBND 데이터로 이동한 것으로 나타났다. Figure 3.51(c), (d)를 보면 inBND 데이터가 outBND 데이터보다 거리가 가까움을 확인하였다.

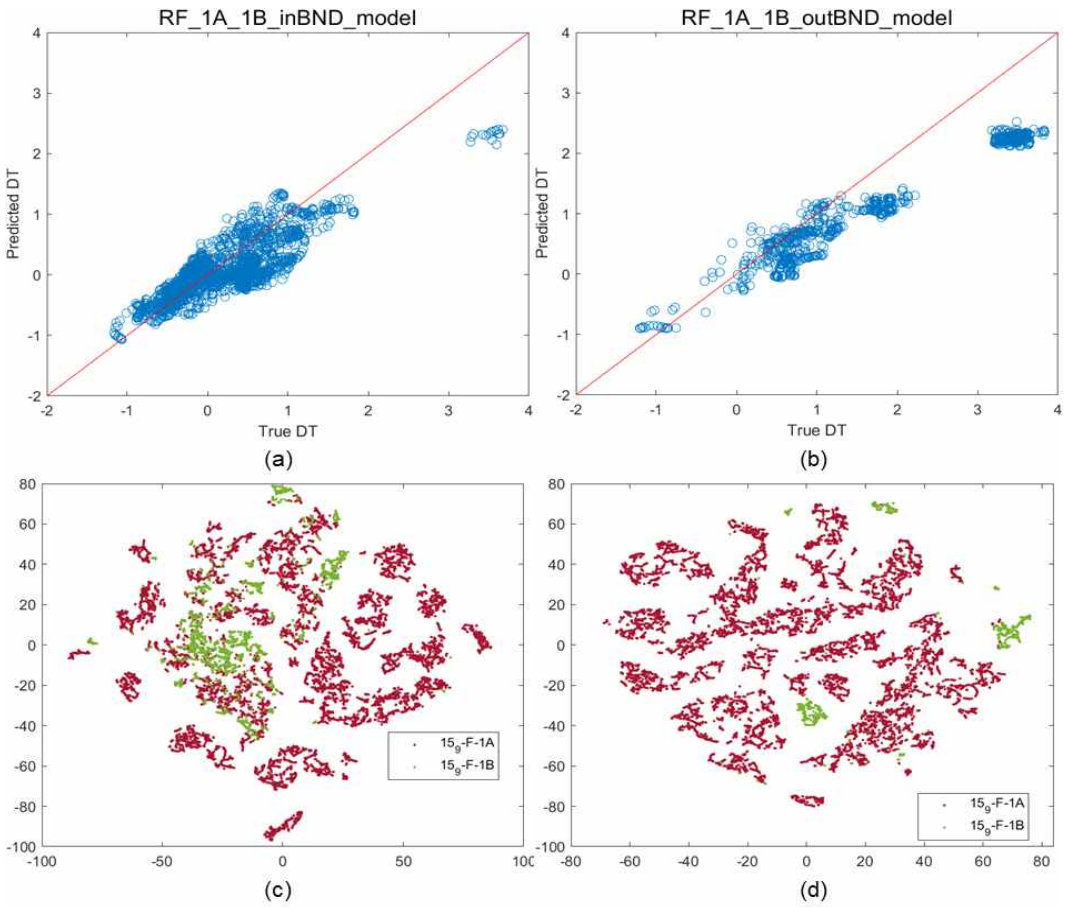


Figure 3.51 Prediction result of 15/9-F-1B using 15/9-F-1A training model for g=1.

나. g=5인 경우

Figure 3.52(a)와 (b)는 15/9-F-1A 데이터를 기준으로 SVDD를 사용해 g=5로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-1B의 DT를 예측한 결과이다. Figure 3.51(a), (b)와 비교했을 때, g를 5로 증가시키자 inBND에서 예측이 잘 맞지 않는 데이터가 outBND로 이동한 것을 확인할 수 있었다.

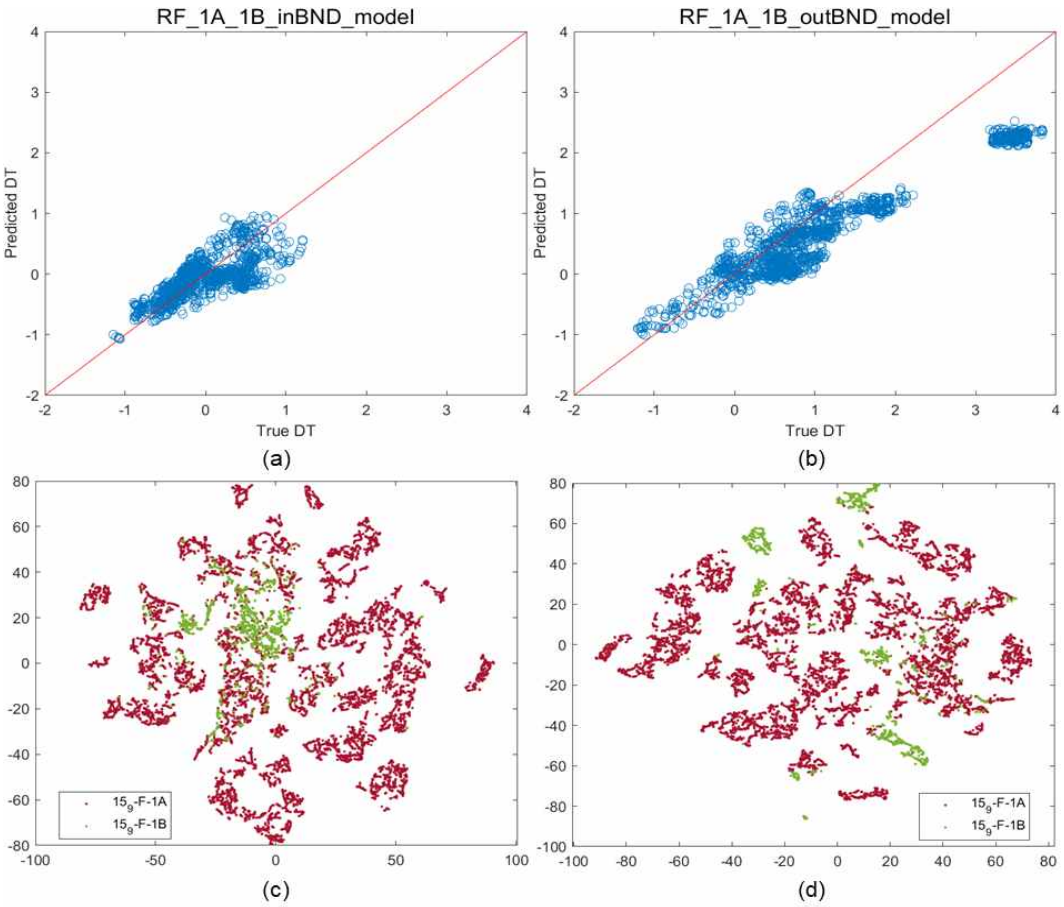


Figure 3.52 Prediction result of 15/9-F-1B using 15/9-F-1A training model for g=5.

Figure 3.53는 15/9-F-1A 데이터를 학습하여 15/9-F-1B의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.5011로 9개의 모델 중 가장 크다.

inBND RMSE는 g=1일 때 0.3509, g=3일 때 0.3151, g=5일 때 0.2987로 오차가 감소하는 것으로 나타났으며, outBND RMSE는 g=1일 때 0.7738, g=3일 때 0.6759, g=5일 때 0.6293으로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다. g값이 증가함에 따라서 inBND와 outBND 오차가 모두 감소하는 경향을 보였다.

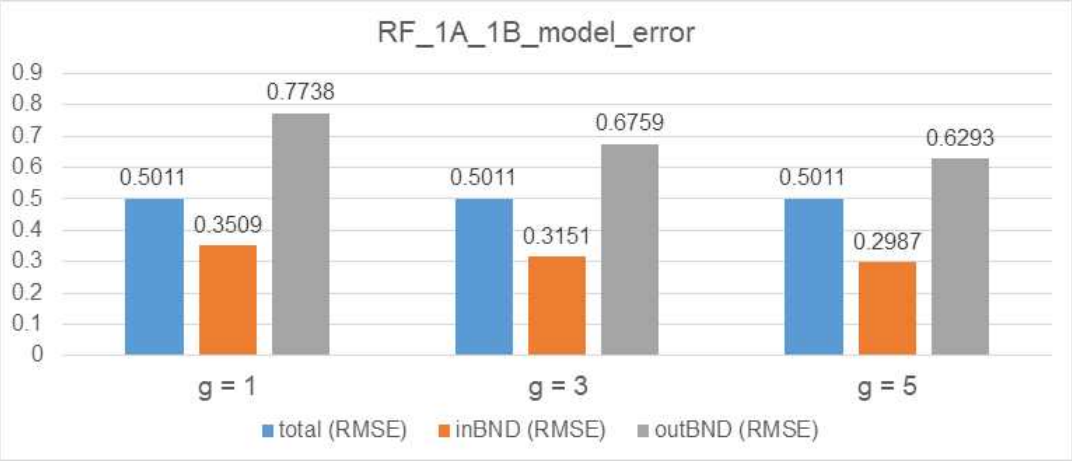


Figure 3.53 Comparison result of RF_1A_1B_model for g value change

Table 3.12는 15/9-F-1A 데이터를 학습하여 15/9-F-1B의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각 각 73.11%, 57.56%, 47.26%이며, 경계 외부에 해당하는 데이터의 비율은 26.89%, 42.44%, 52.74%로 증가하였다.

전체 RMSE 대비 inBND RMSE는 상대적으로 g=1일 때, 29.97%, g=3일 때, 37.12%, g=5일 때, 40.39%로 크게 감소하였으며, outBND RMSE는 g=1일 때, 54.42%, g=3일 때, 34.88%, g=5일 때, 25.58% 증가하였다.

Table 3.12 Relative difference result of RF_1A_1B_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	73.11 %	57.56 %	47.26 %
ratio of data out of boundary	26.89 %	42.44 %	52.74 %
(1)total RMSE	0.5011		
(2)inBND RMSE	0.3509	0.3151	0.2987
(3)outBND RMSE	0.7738	0.6759	0.6293
relative difference between (1) and (2)	-29.97%	-37.12%	-40.39%
relative difference between (1) and (3)	54.42 %	34.88 %	25.58%

Figure 3.54는 15/9-F-1A 데이터를 학습한 모델로 15/9-F-1B의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간 중 시작 지점과 끝 지점은 C, D 영역으로 분포하였고, 중간 구간에 A, B 영역이 나타났다. 두 개의 시추공 데이터를 학습한 모델의 결과인 Figure 3.50과 비교했을 때, 해당 결과의 A, B에 해당하는 영역이 Figure 3.50에서 C, D 영역으로 변화한 것을 확인할 수 있다. 따라서 15/9-F-1A를 단독으로 학습하는 방법보다는 15/9-F-1A와 15/9-F-11A의 데이터를 학습하여 예측한 결과의 신뢰도가 더 높다고 판단하였다.

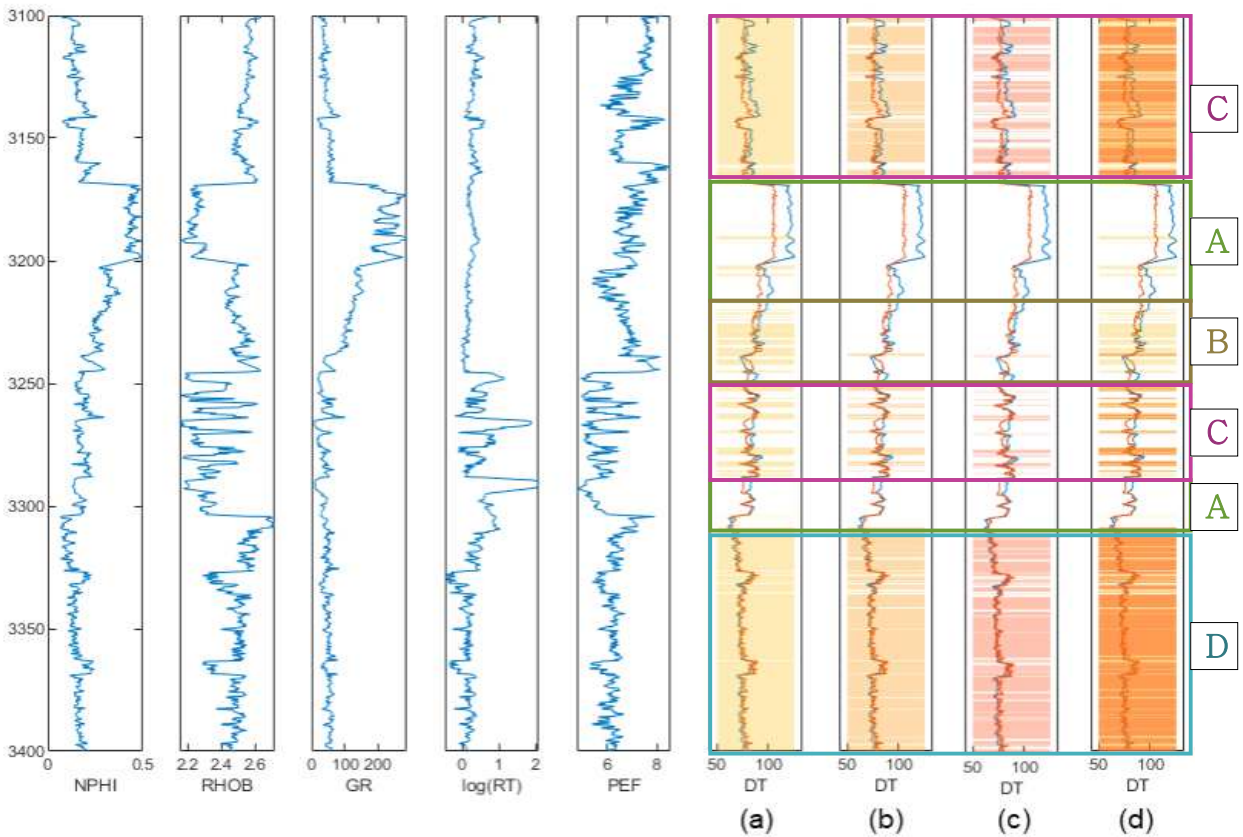


Figure 3.54 Well log of input data and predicted DT of RF_1A_1B_model for g value change

3) 15/9-F-11A를 학습자료로 사용한 경우

가. g=1인 경우

Figure 3.55(a)와 (b)는 15/9-F-11A 데이터를 기준으로 SVDD를 사용해 g=1으로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-1B의 DT를 예측한 결과이다. inBND 데이터에 예측이 잘 맞지 않는 데이터가 포함된 것을 Figure 3.55(a)에서 확인할 수 있다. Figure 3.55(c)에서 거리가 멀리 떨어진 데이터가 inBND로 포함된 것으로 나타났다.

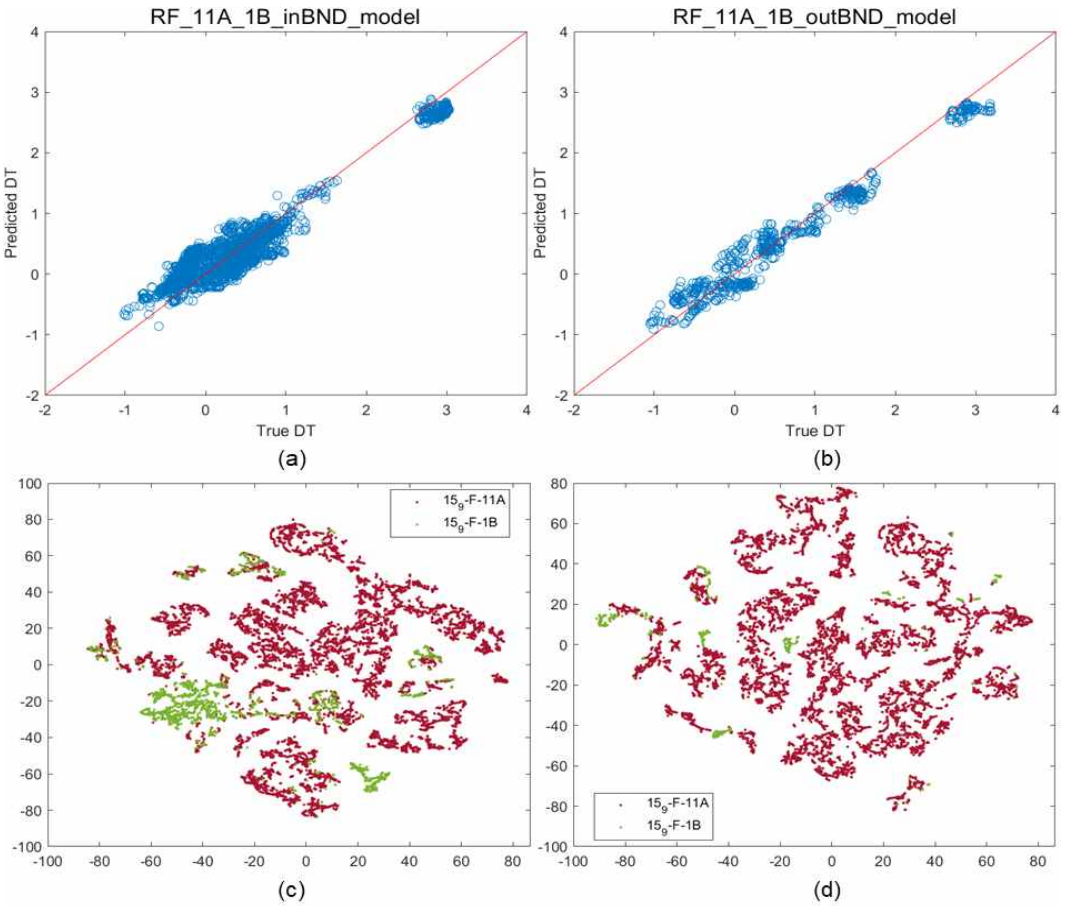


Figure 3.55 Prediction result of 15/9-F-1B using 15/9-F-11A training model for g=1.

나. g=5인 경우

Figure 3.56(a)와 (b)는 15/9-F-11A 데이터를 기준으로 SVDD를 사용해 g=5로 설정하였을 때 생성되는 경계를 바탕으로 내부에 있는 데이터와 외부에 있는 데이터를 나누어 15/9-F-1B의 DT를 예측한 결과이다. g=1인 경우의 결과인 Figure 3.55(a), (b)와 해당 결과를 비교했을 때, inBND에서 예측이 잘 맞지 않는 데이터가 outBND로 이동한 것을 확인할 수 있다. Figure 3.56(c)의 inBND 데이터의 거리 또한 가까운 것으로 보였다.

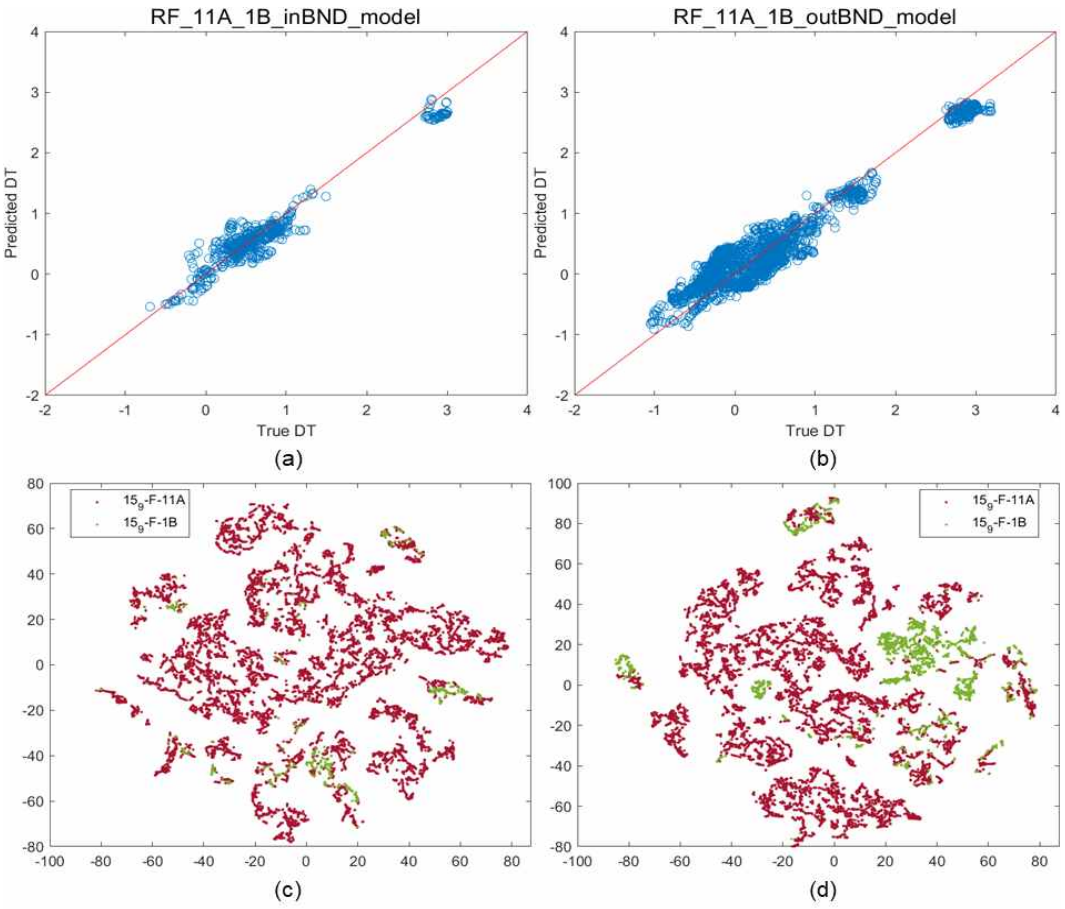


Figure 3.56 Prediction result of 15/9-F-1B using 15/9-F-11A training model for g=5.

Figure 3.57는 15/9-F-11A 데이터를 학습하여 15/9-F-1B의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차이다. 전체 오차는 0.2538이다.

g=1일 때의 inBND RMSE는 0.2631로 전체 오차보다 증가했고, outBND RMSE는 0.2189로 전체 오차보다 감소했다. 해당 결과는 15/9-F-11A 데이터를 대상으로 g=1인 경우로 경계를 생성했을 때 실제로 예측하기 어려운 데이터가 경계 내부에 포함되었을 것으로 판단할 수 있다.

inBND RMSE는 g=3일 때 0.3151, g=5일 때 0.2987로 오차가 감소하는 것으로 나타났으며, outBND RMSE는 g=3일 때 0.6759, g=5일 때 0.6293으로 전체 데이터를 가지고 학습할 때보다는 높은 오차를 보였다. 따라서 g값이 1에서 3으로 증가시켰을 때 앞서 다른 모델들의 결과처럼 전체 오차보다 inBND RMSE가 더 작고, outBND RMSE가 더 큰 결과를 나타냈다.

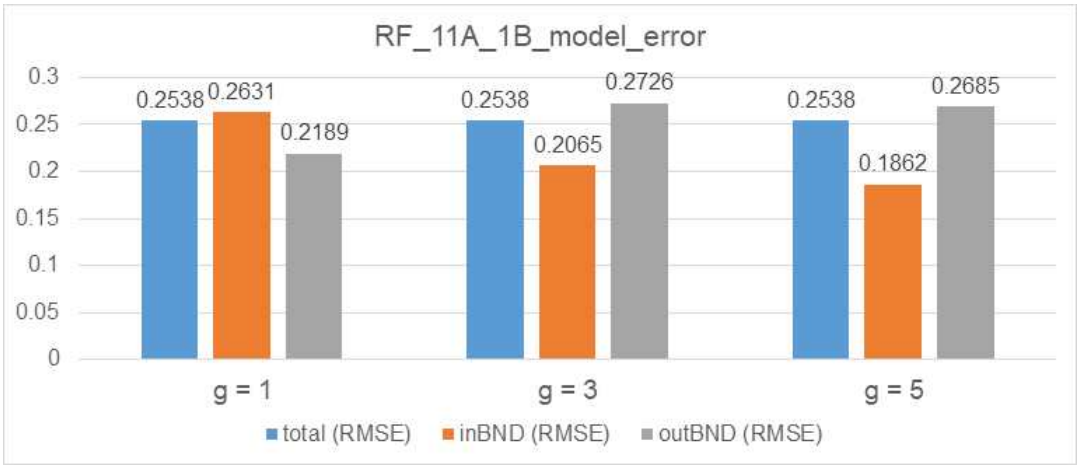


Figure 3.57 Comparison result of RF_11A_1B_model for g value change.

Table 3.13은 15/9-F-11A 데이터를 학습하여 15/9-F-1B의 DT를 예측한 모델을 g 값을 1, 3, 5로 변화시켰을 때의 오차 및 상대 오차율을 나타낸다. g를 1, 3, 5로 증가시켰을 때 설정되는 경계 내부로 포함되는 데이터의 비율은 각 각 77.43%, 31.35%, 20.55%이며, 경계 외부에 해당하는 데이터의 비율은 22.57%, 68.65%, 79.45%로 증가하였다.

전체 RMSE 대비 inBND RMSE는 g=1일 때 3% 증가하고, outBND RMSE는 13.75% 감소했다. inBND RMSE는 g=3일 때, 18.64%, g=5일 때, 26.64%로 감소하

였으며, outBND RMSE는 g=3일 때, 7.41%, g=5일 때, 5.79% 증가하였다.

Table 3.13 Relative difference result of RF_11A_1B_model for g value change

	g = 1	g = 3	g = 5
ratio of data within boundary	77.43 %	31.35 %	20.55 %
ratio of data out of boundary	22.57 %	68.65 %	79.45 %
(1)total RMSE	0.2538		
(2)inBND RMSE	0.2631	0.2065	0.1862
(3)outBND RMSE	0.2189	0.2726	0.2685
relative difference between (1) and (2)	3.66 %	-18.64 %	-26.64 %
relative difference between (1) and (3)	-13.75 %	7.41 %	5.79 %

Figure 3.58는 15/9-F-11A 데이터를 학습한 모델로 15/9-F-1B의 DT를 예측한 결과를 나타낸 그래프이다. 전체 예측 구간 중 절반 이상은 C 영역으로 분포하였고, 일부분 구간은 A 영역, 끝 구간은 B 영역이 나타났다. 두 개의 시추공 데이터를 학습한 모델의 결과인 Figure 3.50과 비교했을 때, 해당 결과의 A, B에 해당하는 영역이 Figure 3.50에서 C, D 영역으로 변화한 것을 확인할 수 있다. 따라서 15/9-F-11A를 단독으로 학습하는 방법보다는 15/9-F-1A와 15/9-F-11A의 데이터를 학습하여 예측한 결과의 신뢰도가 더 높다고 판단하였다.

Table 3.14에는 SVDD 기법을 g값에 따라 9개의 모델에 모두 적용하였을 때, 전체 오차와 inBND 및 outBND의 상대 오차율을 정리한 표이다.

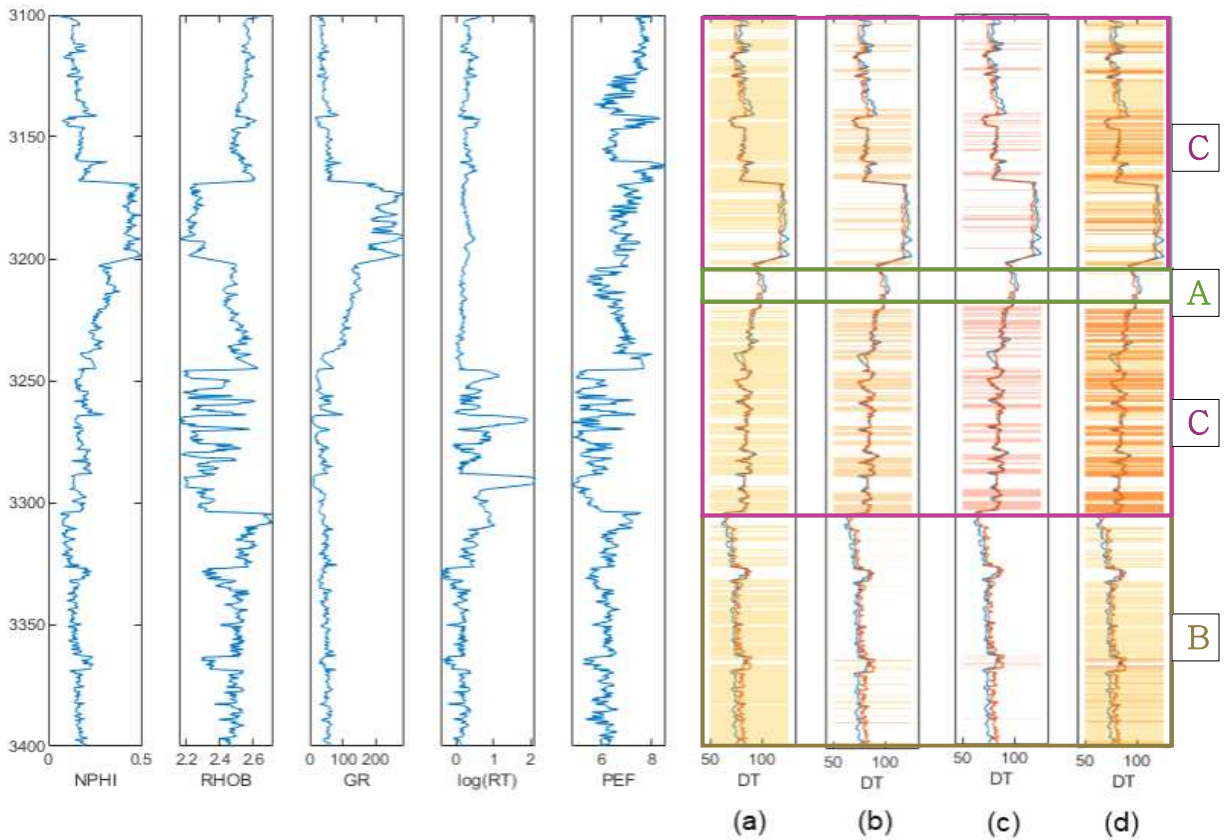


Figure 3.58 Well log of input data and predicted DT of RF_11A_1B_model for g value change

Table 3.14 Relative error of total models for g value change

	g = 1		g = 3		g = 5	
	INBND relative error	OUTBND relative error	INBND relative error	OUTBND relative error	INBND relative error	OUTBND relative error
1A_1B	-29.97%	54.42%	-37.12%	34.88%	-40.39%	25.58%
1A_11A	-16.60%	38.09%	-26.38%	27.20%	-29.53%	20.74%
1B_1A	-37.30%	24.16%	-48.83%	12.96%	-49.56%	7.02%
1B_11A	-56.95%	17.77%	-64.26%	6.98%	-66.51%	3.66%
11A_1A	-11.96%	34.33%	-17.57%	11.14%	-25.09%	7.93%
11A_1B	3.66%	-13.75%	-18.64%	7.41%	-26.64%	5.79%
1A+1B_11A	-7.10%	21.40%	-12.19%	19.34%	-18.79%	16.36%
1A+11A_1B	-2.10%	22.98%	-4.21%	17.80%	-6.23%	12.96%
1B+11A_1A	-17.17%	65.20%	-17.94%	22.65%	-20.31%	13.59%
Mean	-19.50%	29.40%	-27.46%	17.82%	-31.45%	12.63%

제4장 미측정된 음파 검층 예측

제 4장에서는 15/9-F-1A, 15/9-F-1B, 15/9-F-11A 데이터를 모두 학습하여 생성한 랜덤 포레스트 모델을 이용하여 음파 검층이 미측정된 15/9-F-1C, 15/9-F-11B 데이터의 DT를 예측하였다.

1. 15/9-F-1C 예측 결과

Figure 4.1은 15/9-F-1A, 15/9-F-1B, 15/9-F-11A 데이터를 학습한 결과와 oob 데이터를 이용하여 검증한 결과이다. 학습 결과의 RMSE는 0.1135이며, 검증 결과의 RMSE는 0.1708로 모델의 학습이 비교적 잘 된 것을 확인하였다.

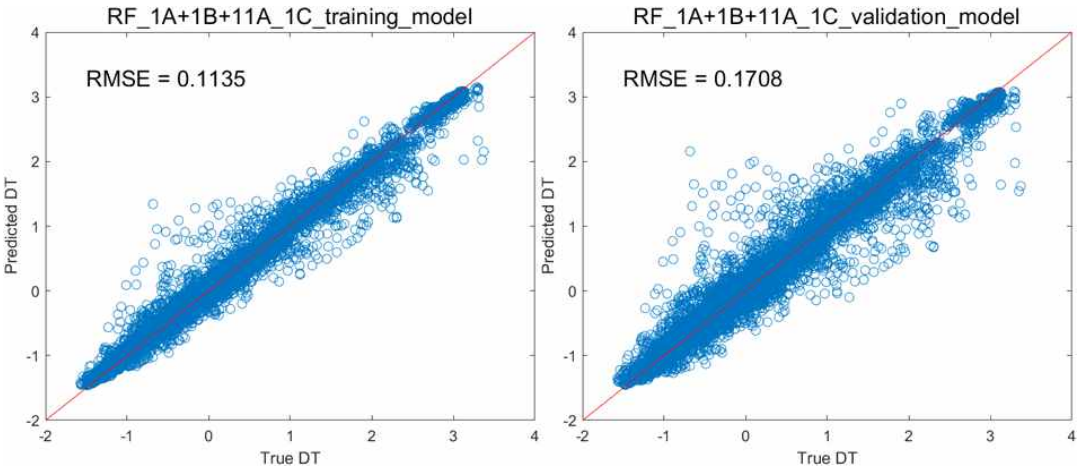


Figure 4.1 Training and validation result of RF_1A+1B+11A_1C_model.

또한 Figure 4.2를 통해서 g가 증가할수록 inBND 데이터에서 거리가 멀리 떨어져 있는 데이터가 outBND 데이터에 다수 포함되는 것을 확인할 수 있었다.

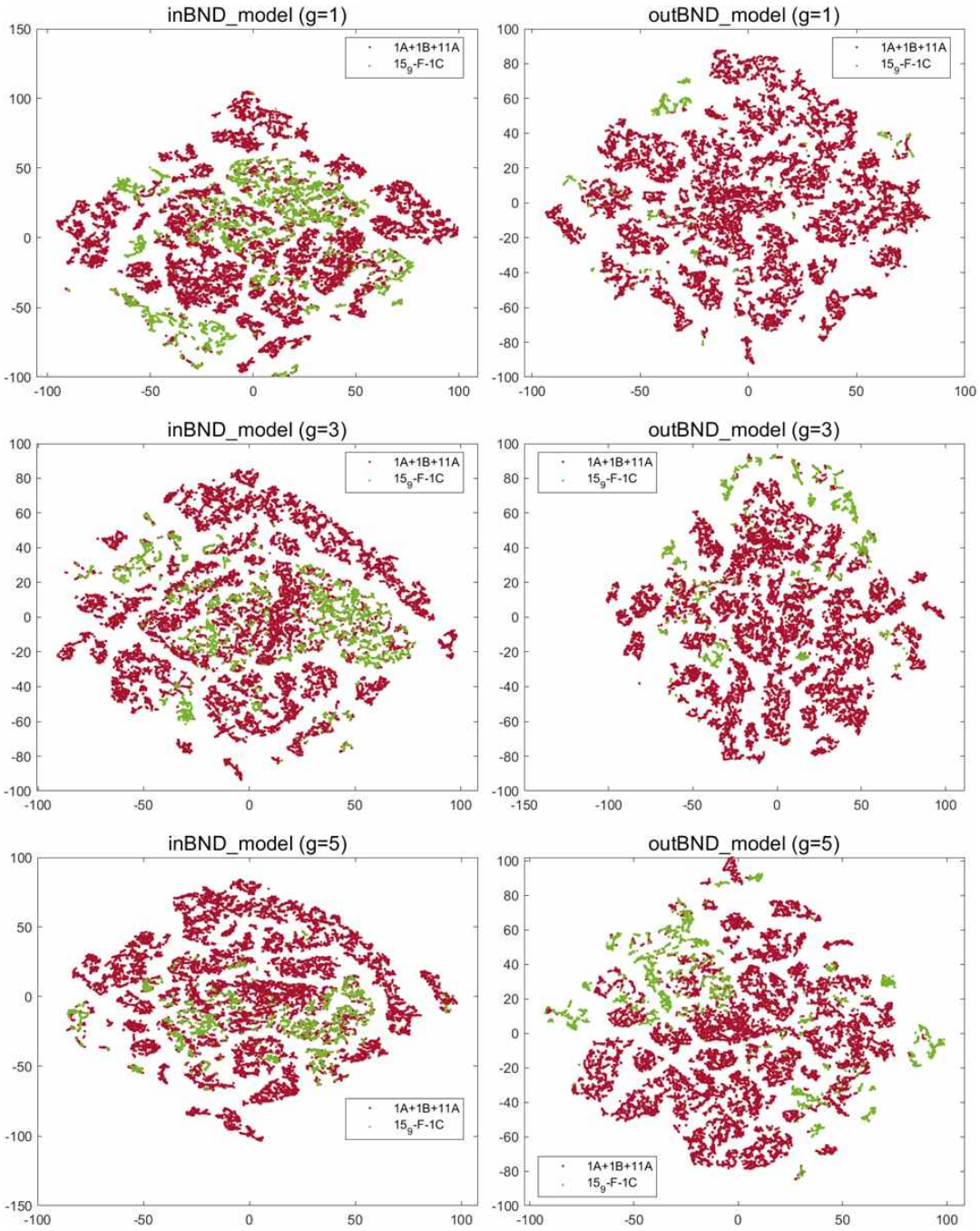


Figure 4.2 Tsne result of RF_1A+1B+11A_1C_model for g value change

Figure 4.3(a)는 g=1일 때, Figure 4.3(b)는 g=3일 때, Figure 4.3(c)는 g=5일 때의 inBND 데이터의 예측 영역이다. Figure 4.3(d)는 세 영역을 모두 합쳐 나타내었다.

Figure 4.3을 분석해본 결과, A영역은 전체에서 일부만을 차지하며 대부분 C와 D영역으로 구분되어 있어 A영역을 제외하고 전체적으로 예측이 잘 되었다고 판단할 수 있다.

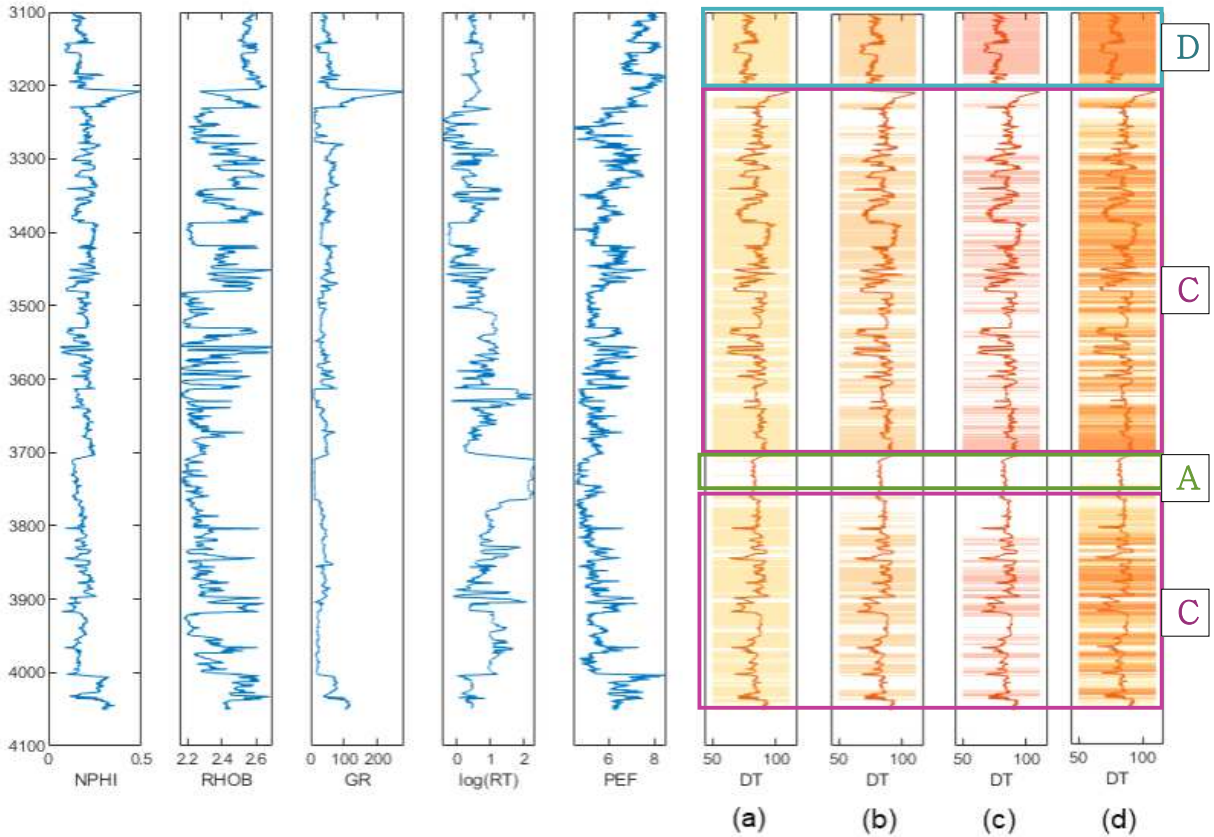


Figure 4.3 Well log of input data and predicted DT of RF_1A+1B+11A_1C_model for g value change.

2. 15/9-F-11B 예측 결과

Figure 4.4는 15/9-F-1A, 15/9-F-1B, 15/9-F-11A 데이터를 학습한 결과와 oob 데이터를 이용하여 검증한 결과이다. 학습 결과의 RMSE는 0.1135이며, 검증 결과의 RMSE는 0.1708로 모델의 학습이 비교적 잘 된 것을 확인하였다. 해당 모델을 사용하여 15/9-F-11B 데이터의 DT를 예측하였다.

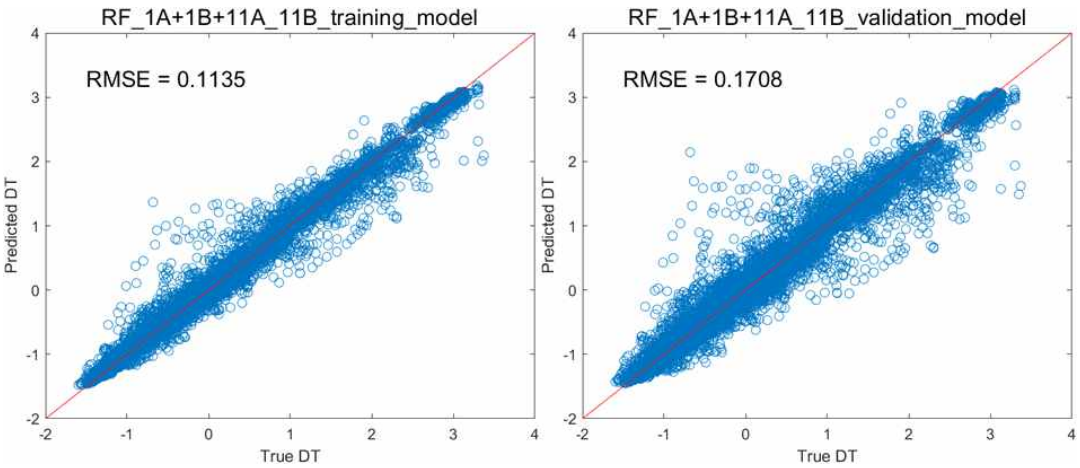


Figure 4.4 Training and validation result of RF_1A+1B+11A_11B_model

또한 Figure 4.5를 통해서 g가 증가할수록 inBND 데이터에서 거리가 멀리 떨어져 있는 데이터가 outBND 데이터에 다수 포함되는 것을 확인할 수 있었으며, 특히 g=1인 경우와 g=5인 경우를 비교해봤을 때 거리 차이를 명확히 판단할 수 있었다.

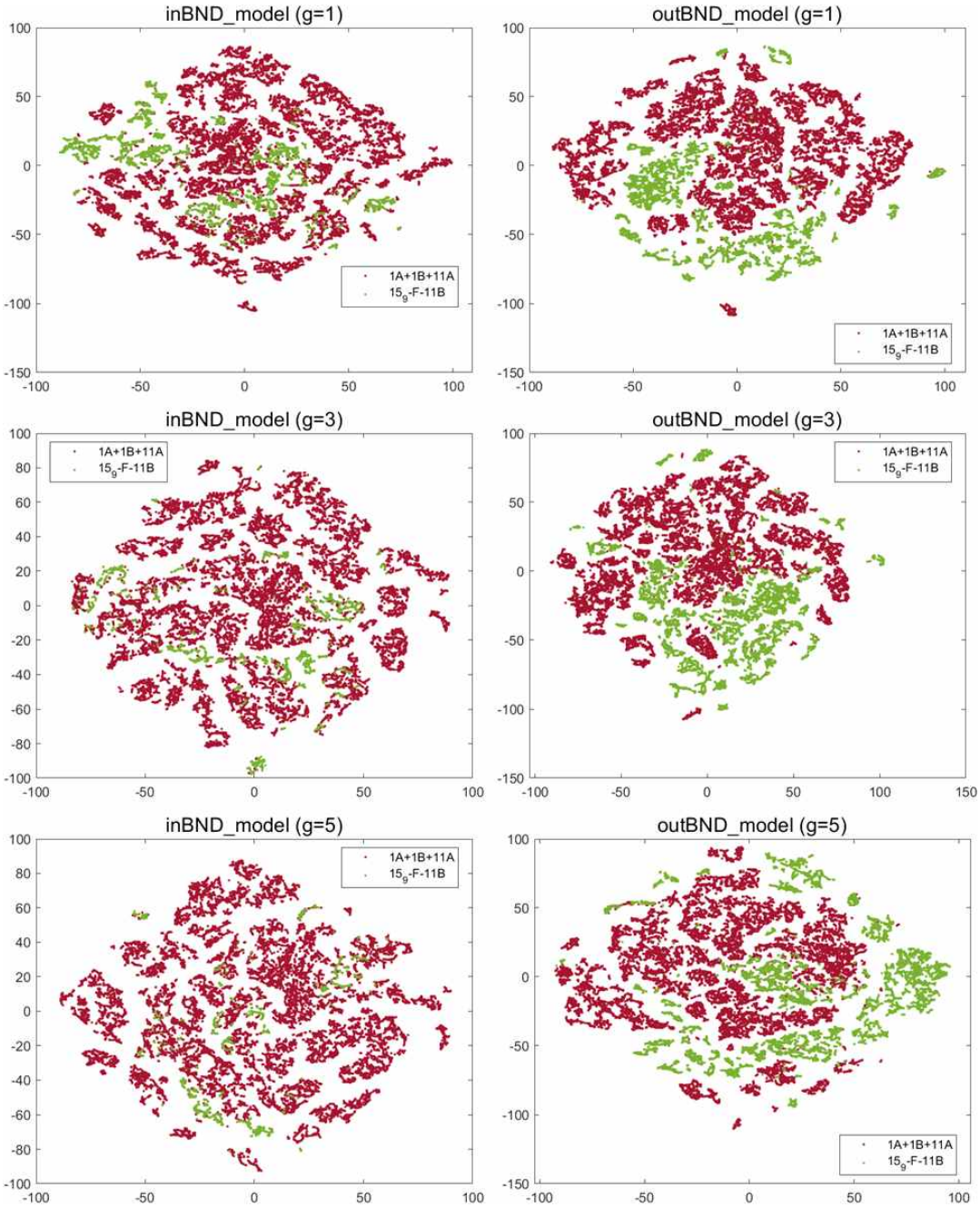


Figure 4.5 Tsne result of RF_1A+1B+11A_11B_model for g value change

Figure 4.6(a)는 g=1일 때, Figure 4.6(b)는 g=3일 때, Figure 4.6(c)는 g=5일 때의 inBND 데이터의 예측 영역이다. Figure 4.6(d)는 세 영역을 모두 합쳐 나타내었다.

Figure 4.6을 분석해본 결과, 전체 예측 구간 중 절반 이상이 A, B영역으로 구분되었으며, C영역은 중간에 일부분, D영역은 초반 영역에 나타났다. 따라서 15/9-F-11B의 DT를 예측한 결과의 신뢰도가 높다고 하기 어려웠으며, 데이터 간 유사성이 낮은 것으로 해석할 수 있다.

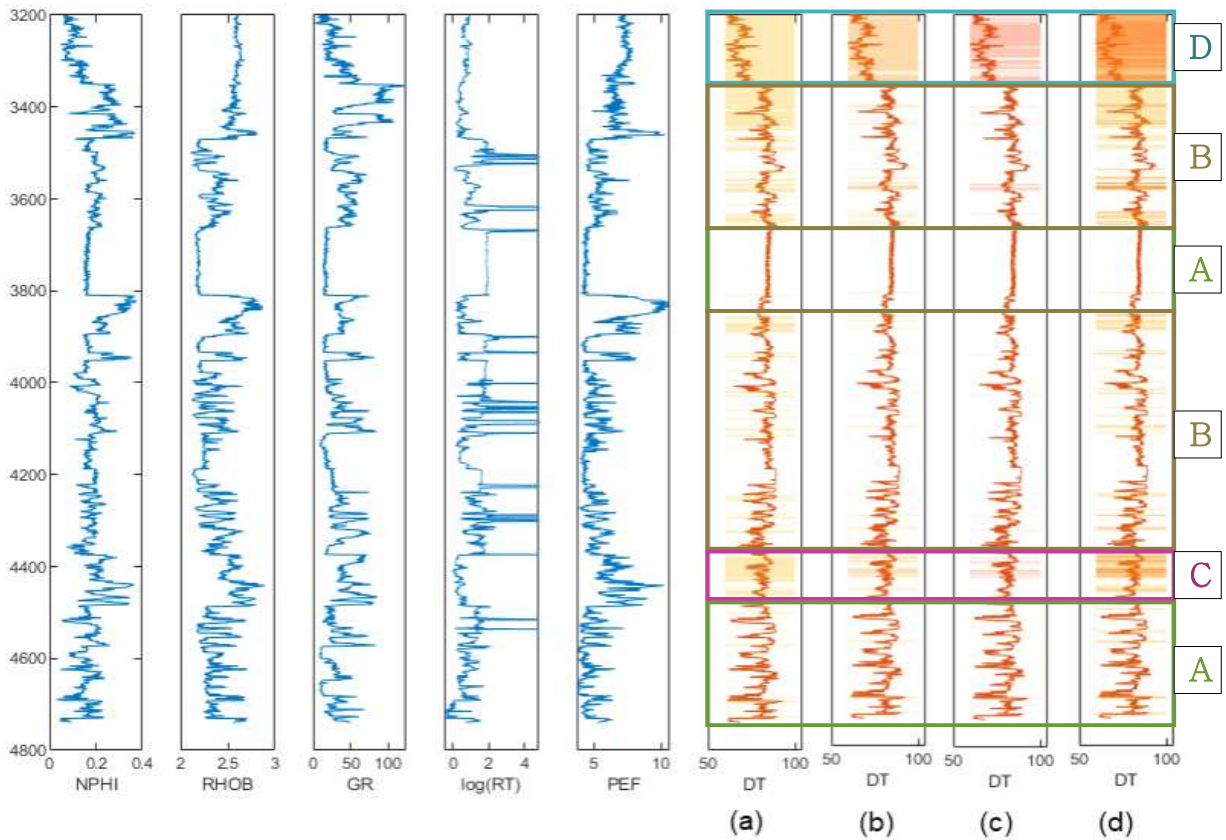


Figure 4.6 Well log of input data and predicted DT of RF_1A+1B+11A_11B_model for g value change

제5장 결론

본 연구에서는 volve 유전의 시추공 물리검층 데이터를 이용하여 미측정된 음파 검층 (DT)를 예측하는 모델에 대한 연구를 수행하였다. 원본 데이터는 결측치를 다수 포함하고 있어 데이터 전처리 과정을 통해 데이터를 가공하였다. 입력 데이터는 측정된 검층 항목 간의 상관계수가 높은 특성들을 선택한 결과, NPHI, RHOB, GR, RT, PEF가 선정되었다. 모델 생성에는 랜덤 포레스트 기법을 사용하였다. DT에 대한 정보를 가지고 있는 15/9-F-1A, 15/9-F-1B, 15/9-F-11A 데이터를 이용하여 모델을 만든 후 서로의 DT 값을 예측하였고, 두 시추공의 데이터를 합쳐 학습한 후 나머지 하나의 데이터의 DT를 예측하여 결과를 비교하였다. 또한 SVDD를 사용하여 g 값 변화에 따라 inBND와 outBND를 구분하여 각각의 오차를 구하여 전체오차와 비교 분석을 수행하였으며, 예측 신뢰도가 높은 구간을 물리검층 그래프에 표시하였다. 이와 같은 방식을 DT가 결여된 15/9-F-1C, 15/9-F-11B에 적용하여 DT를 예측하였으며, 예측 신뢰도가 높을 것으로 추정되는 구간을 선정할 수 있었다.

1. 15/9-F-1C, 15/9-F-11B 데이터는 DT가 결측된 데이터이기 때문에 예측 결과의 검증이 불가능하다. 하나의 시추공 데이터를 학습한 후 다른 데이터의 DT를 예측한 결과의 평균 오차는 0.3453이었으며, 시추공 두 개를 결합한 데이터를 학습하여 예측한 DT의 평균 오차는 0.2358이었다. 따라서 15/9-F-1C, 15/9-F-11B의 DT를 예측하기 위해 15/9-F-1A, 15/9-F-1B, 15/9-F-11A의 데이터를 모두 결합하여 학습한 모델을 이용하면 다양한 입력데이터의 분포를 학습하므로 예측 결과 오차가 가장 작을 것으로 판단된다.

2. SVDD를 사용하여 g값을 1, 3, 5로 증가시켜 inBND 데이터의 오차와 outBND 오차 결과를 분석하였다. 전체 데이터에 비해 inBND 데이터의 오차는 g=1일 때 평균적으로 19.50 %, g=3일 때 27.46 %, g=5일 때 31.45 % 더 낮게 나오는 것으로 분석되었다. 따라서 학습 데이터를 기준으로 가깝게 경계를 설정하여 해당 경계 내부에 포함된 데이터를 예측했을 때 더 좋은 결과를 도출할 수 있음을 확인하였다. 또한 g값을 1, 3, 5로 증가시켜 도출한 inBND 데이터를 물리검층 그래

프 영역에 표시함으로써, 예측 신뢰도가 높은 구간의 파악이 용이하도록 하였다.

3. 예측 신뢰도의 영역은 A, B, C, D 영역으로 구분하였으며, A영역은 g=1, 3, 5인 경우에 관계 없이 대부분 outBND에 속하며, B영역은 g=1일 때, inBND 데이터로 해석되는 데이터가 존재하지만 g=3, 5일 때는 outBND로 데이터가 이동한 것으로 보였다. 그리고 C영역은 g=1일 때, 대부분 inBND 데이터이지만 g=3, 5일 때는 간헐적으로 outBND에 포함되었다. 마지막 D영역은 g=1, 3, 5인 경우 모두 inBND에 포함되는 것으로 보였다. 해당 방법을 통해 미측정된 로그의 전체 예측 구간 중 신뢰도가 높은 영역과 낮은 영역을 구분하여 제시할 수 있다.

이 연구에서 제시한 방법론은 미측정 검증 자료를 예측할 때 예측 결과의 신뢰도를 분석할 수 있어 높은 신뢰도를 나타내는 구간을 확인하는 방법으로 활용도가 높을 것으로 판단된다. 또한 해당 방법론은 물리 검증 자료를 예측하는 소프트웨어를 개발할 경우에도 예측 신뢰도를 제시해 주는 유용한 기능이 될 것으로 예상된다.

참고문헌

김성필, 2016, “딥러닝 첫걸음”, 한빛미디어(주), pp. 17-26.

김은미, 2020, 국내 주택시장의 주택 보유기간 및 매도 의사결정에 대한 머신러닝 예측모델 비교, 석사학위논문.

김준석, 2021, 공간랜덤포레스트의 버퍼거리를 이용한 월 평균기온, 월 누적 강수량 예측성능 비교, 석사학위논문.

박해선, 2021, “핸즈온 머신러닝”, 한빛미디어(주), pp. 246-257.

안성호, 2017, SVDD를 활용한 다중학습 기반 랜섬웨어 탐지 방법에 관한 연구, 박사학위논문.

오현택, 2020, 유가스정 생산추이 예측을 위한 머신러닝 모델 구축 연구, 석사학위논문.

이예지, 2020, 랜덤포레스트 모델링기법을 이용한 한강유역 지류의 저서동물지수 예측에 관한 연구, 석사학위논문.

임성태, 2019, 비지도학습 머신러닝 기반 자기주도학습 역량 잠재요인의 군집분석 모델 개발, 석사학위논문.

Alizadeh, B., Najjari, S., and Kadkhodaie-Ilkhchi A., 2012, Artificial neural network modeling and cluster anaylsis for organic facies and burial history estimation using well log data:a case study of the South Pars Gas Field, Persian Gulf, Iran, Computers & Geosciences, 45, p.261-269.

B. Scholkopf, R. C. Williamson, A. J. Smola, J. S. Taylor, and J.C. Platt. 2000, Support vector method for novelty detection. In Neural Information Processing Systems, pages, 582-588.

Bailey, T., Dutton, D., 2012, An Empiricial Vp/Vs Shale Trend for the Kimmeridge Clay of the Central North Sea.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., 1984, Classification and regression trees. Wadsworth CA, USA.

Breiman, L., 1994, Bagging Predictors, Machine Learning, Kluwer Academic Publishers. Boston. 24, 123-140.

Breiman, L., 2001, Random forest, Machine Learning, Kluwer Academic Publishers, Netherlands, 45(1), pp. 5-32.

Equinor, 2018. Volve data village. https://data.equinor.com/dataset/Volve.

Feng, R., Grana, D., Balling, N., 2021, Imputation of missing well log data by random forest and its uncertainty anaylsis. Computers and Geosciences. 152, 104763.

Geron, A., 2019, Hands-on Machine Learning with Scikit-learn, Keras & Tensorflow, O'reilly, Sebastopol, CA, 260p.

Hossain, Z., Mukerji, T., Fabricius, I.L., 2012. Vp-Vs relationship and amplitude variation with offset modelling of glauconitic greensand. Geophys. Prospect. 60, 117-317.

Jian, H., Chenghui, L., Zhimin, C., Haiwei, M., 2020, Integration of deep neural networks and ensemble learning machines for missing well logs estimation.

Flow Measurement and Instrumentation. 73, 101748.

Matten, L., Hinton, G., 2008, Visualizing Data using t-SNE, Jounal of Machine Learning Research 1, 1-48.

Miah, M.I., Ahmed, S., Zendehboudi, S., 2021, Model development for shear sonic velocity using geophysical log data:Sensitivity analysis and statistical assessment, Journal of Petroleum Science and Engineering, 88, 103778.

Miller, S., Stewart, R., 1990, Effects of lithology, porosity and shaliness on the P ans S-wave velocities from sonic logs. J. Can. Soc. Explor. Geophys. 26, 94-103.

Onalo, D., Adedigba, S., Khan, F., James, L.A., Butt, S.D., 2018a, Data Driven Model for Sonic Well Log Prediction, J. Petrol. Sci. Eng. 170, 1022-1037.

Otchere, D. A., Arbi Ganat T. O., Ojero, J. O., Tackie-Otoo, B. N., Taki, M. Y., 2021, Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. Journal of Petroleum Science and Engineering. 208, 109244.

Raymer, L.L., Hunt, E.R., Gardner, J.S., 1980, An Improved Sonic Transit Time-to-Porosity Transform. In: SPWLA Annu. Logging Symp, pp. 1-13.

Scholkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J. 1999, Single-class Support Vector Machines. Poster presented at Dagstuhl-Seminar 99121:Unsupervised Learning, Dagstuhl, Germany.

Tax, D., & Duin, R. 2004, Support Vector Data Description, Machine Learning, 54, p.45-66.

Walls, J., Dvorkin, J., Mavko, G., Nur, A., 2000, Use of compressional and shear wave velocity for overpressure detection. In: Pap. OTC 11912 Proc. 2000 Offshore Technol.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. 2016, Data mining:Practical machine learning tools and techniques. Morgan Kaufmann.

Zerrouki, A., Aifa, T., and Baddari, K., 2014, Prediction of natural fracture porosity from well log data by means of fuzzy ranking and an artificial neural network in Hassi Messaoud oil field, Algeria, Journal of Petroleum Science and Engineering, 115(1), p.78-89.

Zeng, L., Ren, W., Shan, L., Huo, F., 2021, Well logging prediction and uncertainty analysis based on recurrent neural network with attention mechanism and Bayesian theory. Journal of Petroleum Science and Engineering. 208, 109458.