energies

MDPI

Article

Probabilistic Prediction of Multi-Wells Production Based on Production Characteristics Analysis Using Key Factors in Shale Formations

Hyo-Jin Shin [1], Jong-Se Lim [1,*] and Il-Sik Jang [2]

[1] Department of Energy & Resources Engineering, Korea Maritime & Ocean University, Busan 49112, Korea; hjshin@kmou.ac.kr
[2] Department of Energy Resources Engineering, Chosun University, Gwangju 61452, Korea; isjang77@Chosun.ac.kr
* Correspondence: jslim@kmou.ac.kr; Tel.: +82-51-410-4682

Abstract: In this study, we propose a novel workflow to predict the production of existing and new multi-wells. To perform reliable production forecasting on heterogeneous shale formations, the features of these formations must be analyzed by classifying the formations into various groups; the groups have different production characteristics depending on the key factors that affect the shale formation. In addition, the limited data obtained from nearby existing multi-wells should be used to estimate the production of new wells. The key factors that affect shale formation were derived from the correlation and principal component analysis of available production-related attributes. The production of existing wells was estimated by classifying them into groups based on their production characteristics. These classified groups also identified the relationship between hydraulic fracturing design factors and productivity. To estimate the production of new wells (blind wells), we generated groups with different production characteristics and leveraged their features to estimate the production. Probabilistic values of the group features were entered into the input layer of the artificial neural network model to consider the variation in the production of shale formations. All the estimated productions exhibited less error than the previous analytical results, suggesting the utilization potential of the proposed workflow.

Keywords: shale formation; probabilistic prediction; production characteristics; key factors; multi-wells

Citation: Shin, H.-J.; Lim, J.-S.; Jang, I.-S. Probabilistic Prediction of Multi-Wells Production Based on Production Characteristics Analysis Using Key Factors in Shale Formations. *Energies* **2021**, *14*, 5226. https://doi.org/10.3390/en14175226

Academic Editors: Reza Rezaee and Yuichi Sugai

Received: 28 July 2021
Accepted: 19 August 2021
Published: 24 August 2021

1. Introduction

In the ever-increasing global demand for energy, unconventional resources now account for a large portion of untapped reserves. Among them, tight/shale gas and oil are expected to be actively produced [1]. In addition, production techniques, such as horizontal drilling and hydraulic fracturing for large-scale production, have been developed for continuously distributed shale formations. The application of hydraulic fracturing for efficient production and the number of horizontal wells is expected to increase. These improvements in the production technology have led to low development costs and high commercial success rates. Moreover, the break-even prices of oil and gas production, which provide equal revenues and costs over a certain period of time from shale formation, are gradually decreasing due to the improvements in production technologies.

However, the heterogeneity of the shale reservoirs with a permeability distribution in the nano-Darcy scale and complex flow mechanisms makes it difficult to use conventional techniques for production forecasting. Conventional analytical methods, which are applied to predict production, have uncertainties and limitations due to subjective judgments and assumptions of input settings and forecasting production trends. Therefore, a reliable method for production forecasting that considers the unclear flow mechanisms during production and native and hydraulic design parameters of the shale formations is required.

The petroleum exploration and production (E&P) industry is currently undergoing a revolution owing to the use of data-driven analytics, which has become an important competitive technology. Regarding production from unconventional resources, shale companies have collected large amounts of data over the last few years, which can be used to optimize production forecasts, exploration risks, production development scenarios, drilling and completion, and production operations.

Recently, data-driven analytics have been employed to predict the performance of shale formations using production-related attributes, such as native, design, and dynamic parameters. However, the complex effects of the attributes do not exhibit specific trends in productivity, and the relations among attributes have not been clearly examined in the existing literature. Therefore, data-driven analytics techniques, such as artificial intelligence and machine learning, were applied to classify production wells and estimate production forecasts in this study. Several studies have shown that data-driven analytics are effective in estimating reservoir properties, predicting production, optimizing drilling, and solving various other issues in shale formations. The results of the decline curve analysis (DCA) and artificial neural networks (ANN) were compared and analyzed to predict the production over time for a single well [2,3], while actual production was predicted using a recurrent neural network [4,5].

However, shale strata that are heterogeneously distributed over a large area require analysis of multiple wells rather than a single well. Mohaghegh [6] identified the correlation among reservoir characteristics, rock-mechanical properties, hydraulic fracturing design factors, and production in shale formations, and evaluated their impact using pattern recognition. In addition, the effects of each key factor and productivity were identified based on field data obtained from the Permian Basin [7], and machine learning was applied to estimate the decline curve factors; estimated ultimate recovery (EUR) was also obtained for production forecasting [8–12]. Furthermore, machine learning has been applied to evaluate the mechanical properties of hydraulic fracturing design factors [13,14]. Machine learning has also been applied for new wells with input factors similar to those of the existing wells to predict the production behavior and cumulative production in these new wells for over six months [15]. Amr et al. [16] developed a prediction model using the data from nearby production wells owing to limited information on new regions; however, despite considering the reservoir characteristics, the influence of heterogeneous shale formations on production of multi-wells was unclear.

Therefore, in this study, we proposed a workflow to predict the production of existing and new multi-wells in shale formations. First, we derived production-related attributes of existing production wells and identified the key factors from the correlated response variables. Accordingly, we identified the relationship between production and key factors and improve the reliability of production forecasts by classifying formations with different production characteristics into groups to estimate the cumulative production of existing wells based on the trend line of each group. For the new wells, we developed a model that could classify groups with limited input attributes, such as key factors, and utilize probabilistic values from these classified groups to perform cumulative production forecasting considering uncertainty using ANN (Figure 1).

Section 1 covers the background and objectives of the study and describes the production-related attributes used for data-driven analytics in the study area. In Section 2, the analysis methods used in shale formations to predict production and recent research trends are explained. Section 3 covers the analytical methods used in the proposed workflow, and the results for existing and new wells obtained from the workflow are summarized in Sections 4 and 5.

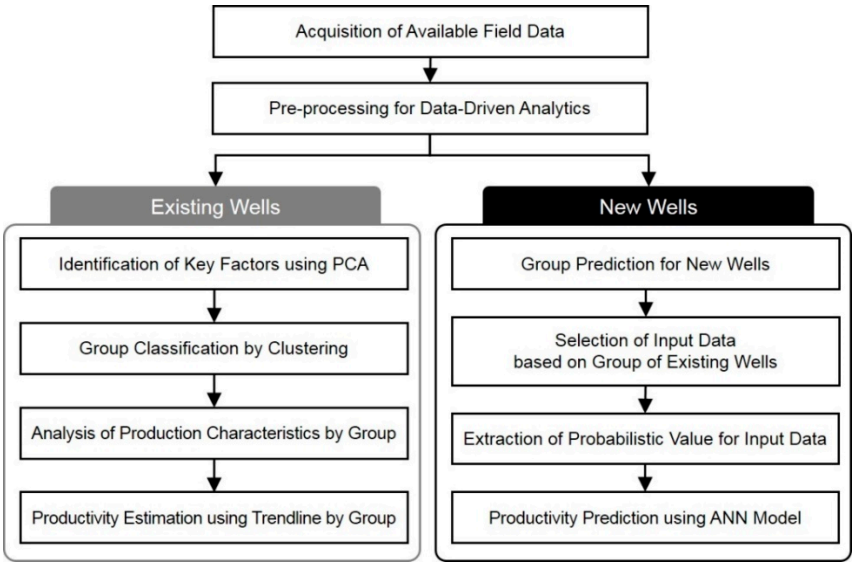


Figure 1. Workflows for predicting productivity from existing and new wells.

1.1. Production-Related Attributes

Production in shale formations is influenced by both the reservoir characteristics and hydraulic fracturing design; it is particularly dependent on well completion technology. The success of hydraulic fracturing indicates that the stimulated reservoir volume (SRV) related to productivity may not be generated homogeneously, or a nonfractured section may persist. By monitoring these shale reservoirs, micro seismic data can be used to determine cracked and unstimulated areas [17], and cracks can be characterized using micro seismic data, production logging, image logging, and tracer tests [18]. However, these methods are expensive in terms of data acquisition or are limited in their ability to individually identify one planar crack in a complex fracture network [19].

Hydraulic fracturing creates artificial cracks by injecting high-pressure fluid blended with proppant into the rock formation [20]. With the increase in diameter of the proppant particles, the conductivity and strength increased. In the case of a fracturing fluid, suitable fluids should be selected based on their sensitivity to formation, pH, and fluid behavior and stability linked to temperature [21]. This results in high initial production and declining rates, leading to a large decline in production in the first five years, with subsequent low production levels. In addition, there are several key challenges in the prediction of production from shale formations, such as a short production period of a significant number of wells, productivity of wells dependent on hydraulic fracturing factors, and uncertain fluid flow mechanisms [2].

Operators have increased the lateral length, number of stages, and amount of hydraulic fracturing fluids on unconventional well completion [22]. For Marcellus shale and Eagle Ford shale, the number of stages increased from 10 in 2008 to 16 in 2012 [23]; for Bakken shale, the number of stages increased from 20 in 2008 to 94 in 2019 [24]. The length of the horizontal wells is usually long, 5000–10,000 ft, and a hydraulic fracturing operation can have as many as 30–40 fracturing stages [25]. For an ideal hydraulic fracturing design, it should be kept in mind that the cost of well completion increases with the number of stages. Thus, further research is required to determine an appropriate hydraulic fracturing design by optimizing the number of stages besides other factors. Field data related to shale production are summarized in Figure 2. However, the relationships between these factors and their effects on production remain unclear [15]. Production changes owing to the complex effects of multiple factors. Therefore, the impact of these factors on productivity should first be identified based on the statistical analysis of each attribute.

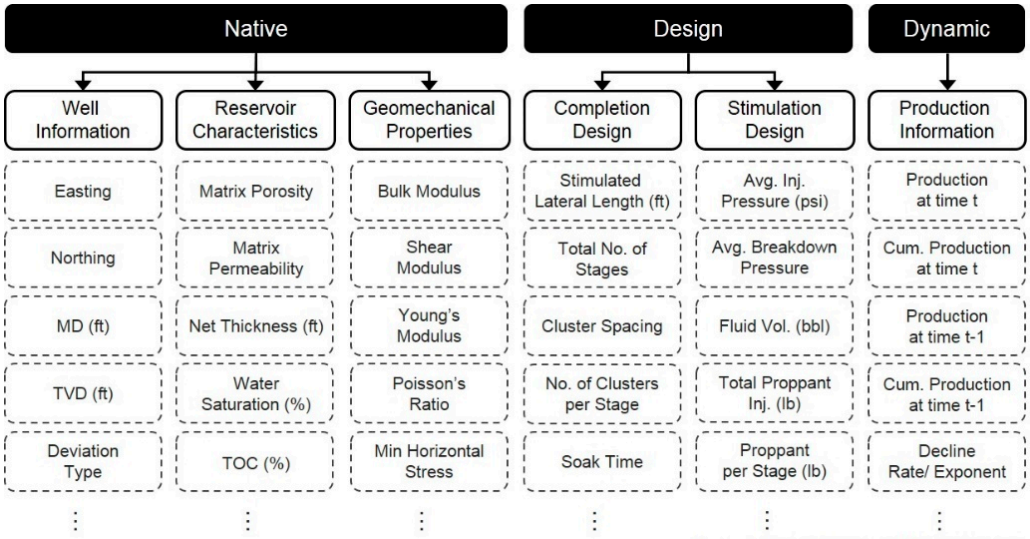


Figure 2. List of native, design, and dynamic attributes in shale formations.

1.2. Study Area

Eagle Ford Shale, one of the most active shale formations in the United States, has more than 100 rigs in operation. The region that forms the source rock for Austin Chalk is estimated to contain 20.81 Tcf of natural gas and 3.351 bbl of oil, wherein various phases of fluids, such as dry gas, wet gas, natural gas liquids, gas condensate, and crude oil are produced. The shale formations have a width and an average thickness of 50 and 250 ft, respectively, and are located at a depth of 4000–14,000 ft [26].

Production data for Eagle Ford shale were acquired from Enverus for the Texas and Louisiana Gulf Coast Basin. The data included the operator, period of production, and duration of well completion; 250 wells were produced in approximately 60 months. The production rate data are generally allocated by commercial data vendors (Enverus or IHS Markit) and are not measured by individual production well gauges. In this study, the oil production data for the target area, that is, the north–west trend area in Eagle Ford Shale, were used, as shown in Figure 3. MATLAB (R2019b, MathWorks) was used for this purpose.

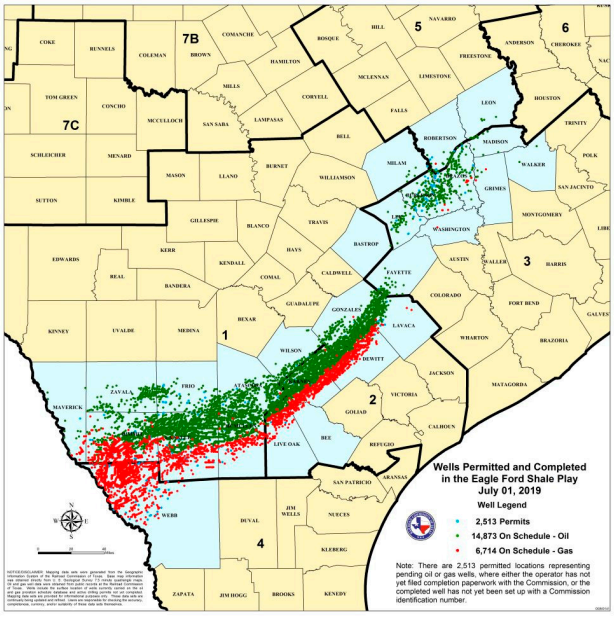


Figure 3. Wells permitted and completed in the Eagle Ford shale [27].

2. Production Estimation Method in Shale Formations

Conventional methods, such as reservoir simulation, rate transient analysis, and DCA, can be used to predict production in shale formations. However, as mentioned earlier, the conventional methods have uncertainties and limitations owing to the subjective judgments and assumptions made in defining input parameters and predicting production trends. Consequently, data-driven analytics, such as artificial intelligence and machine learning, can be used to overcome the limitations of conventional methods and obtain meaningful insights from the results, assuming that meaningful input factors are utilized.

2.1. Decline Curve Analysis

DCA was applied in this study to identify the decline behavior of the curve and predict productivity using only production rate data. If there is no significant physical change around the well, the production changes gradually according to the state of the reservoir. Accordingly, extrapolation can be effectively used to estimate the amount of remaining oil resources and production period. The Arps [28] equation, which is commonly used to predict oil and gas production in the field, can be categorized into three equations according to the decline exponent, namely, exponential, hyperbolic, and harmonic decline curve equations. Although conventional oil and gas resources have a decline exponent in the range 0–1.0, a hyperbolic function is also used in the shale reservoirs with a decline exponent exceeding 1, resulting in an overestimation of EUR. To overcome this problem, various techniques, such as the modified hyperbolic [29], the logistic growth model (LGM) [30], modified stretched exponential production decline (SEPD) [31], and Duong [32] methods, have been proposed. DCA should be performed meticulously because of limitations and subjective judgments regarding its use [33]. Therefore, a suitable DCA differs depending on the type of fluid and the application of production technologies.

The Arps equation, which is commonly used among conventional methods, was used for comparison with the results of the newly applied method in this study. The hyperbolic equation used to predict the production of shale formations in this study is as follows:

$$q_t = q_i(1 + bD_i t)^{-1/b} \quad (1)$$

where q_t is the production rate at time t, q_i is the initial production rate, b is the decline exponent, D_i is the initial decline rate, and t is the time.

The characteristics of the change in the decline curve factor are that the larger the initial decline rate, the steeper the decline is in the initial production; moreover, the larger the decline exponent, the slower is the decrease in the latter half of production [34]. Because shale formation produces oil or gas through cracks generated by hydraulic fracturing, the decline exponent is estimated to be different depending on the period of the production data used [35]. In addition, if there is insufficient production rate data at the beginning of the production, the production forecast results exhibit uncertainty even when using techniques other than DCA [36].

2.2. Probabilistic Method

In shale formations, as the transitional flow is long, boundary-dominated flow may not appear during the production, and the long-term production data may not be sufficient for analysis [37]. Because it is difficult to measure the reservoir properties required when estimating the number of resources using the volumetric method, the Monte Carlo simulation is applied to consider the uncertainty of the number of resources [23].

It is difficult to consider production variability in shale formations when using DCA, which predicts a single value [38]. Therefore, when DCA was applied, a study for calculating the EUR of shale gas was performed using a probabilistic method to predict the number of resources [39,40]. However, when there are limited data, distorted data, etc., there is a limit to applying a histogram in which the shape of the probability density function varies depending on the class intervals or starting points. Therefore, when applying Monte

Carlo simulation by identifying the appropriate distribution form for the decline curve factor using kernel density functions, which is a non-parametric method that analyzes statistics without presupposing inaccurate assumptions, the EUR can be calculated from the probability distribution reflecting the characteristics of the data [41].

2.3. Research Trends

Shale strata that are heterogeneously distributed over a large area require analysis of multiple production wells rather than a single well. Unlike conventional resources, the characteristics of shale reservoirs have not been identified yet. Therefore, many studies that reflect data features over the past three to four years have been conducted using machine learning. The goal of these studies is to predict future production changes for existing wells and forecast the production of a newly drilled well. Further, data-driven analysis should be used to identify unclear correlations between factors affecting the production behavior of shale formations and use them to predict future performance. In this section, we describe the analysis of the research that performed predictions on production of shale formations using machine learning that reflected the characteristics of the shale reservoir; the primary contents are summarized in Tables 1 and 2.

Table 1. Research on productivity prediction of existing wells in shale formations.

Author(s)	**[9]**	**[10]**	**[11]**	**[12]**	**[42]**
Basin	Reservoir simulation	Marcellus	Eagle Ford	Duvernay	Bakken
Fluid	Oil & Gas	Gas	Oil	Oil	Oil & Gas
Well Num.	100	100	120	262	2061
Method	ANN [7 50 3]	ANN [20 20 1]	RF SVM MARS	LR Multiple LR ANN	Deep learning [8 4*100 1]
Input	7	20	9	21	8
	• Permeability • Porosity • Fracture width • Fracture half length • Fracture conductivity • Formation pressure/temperature	• Total proppant • Lateral length • MD, stages • Fracture clusters • Avg. pressure • Easting northing • TOC • Azimuth • Other	• Latitude • Longitude • Initial prod. • Total proppant • Total fluid • Stages • Lateral length • TVD of heel • TVD heel–toe difference	• Azimuth • TVD • Total proppant • Total fluid • Lateral length • Avg. pump rate • Fracture spacing • Fracture stages • Acid volume • Other	• Total thickness • Norm proppant • Depth • Porosity • Stages • Norm fluid • Norm spacing • Water saturation
Output	LGM (K, a, n)	50 years EUR	DCA parameters & EUR (Arps, SEPD. Duong, Weibull)	Well performance (individual well/type curve)	1-year production by stage
Results	R 0.92	The points of the actual and predicted data are almost identical	RMSE (for EUR) 45,32,44,30 (RF) 44,31,43,29 (SVM) - (MARS)	R^2 0.41 (LR) 0.76 (multi LR) 0.93 (ANN)	R^2 0.75 (training) 0.61 (testing)

Table 2. Research on productivity prediction of new wells in shale formations.

Author(s)	**[15]**	**[16]**
Basin	Marcellus	DJ, Williston, Anadarko, Powder River
Fluid	Oil & Gas	Oil
Well no.	128	713 (DJ)
Method	ANN [9 15 1]	Extreme gradient boosting tree (xgbTree) (Best performing of 7 algorithms)
Input	9	31
	• TVD • Net thickness • Porosity • TOC • Lateral length • Total no. of stage • No. of clusters per stage • Fluid per lateral • Proppant per lateral	• Surface/bottom latitude & longitude • County • Lateral length • Proppant lateral • 3, 6, 12 Cum. oil production • Reservoir properties (bulk volume, thickness, porosity, water saturation) • Distance, angle (neighbor wells) • Avg. 3, 6, 12 Cum. oil production (neighbor wells) • Avg. reservoir properties (neighbor wells) • Etc.
Output	6-month Cum. production	Arps (q, D_i, EUR) (one of each parameter)
Results	R^2 0.96 (training) 0.77 (testing)	Accuracy (%) (for best case) 97.31, 87.82, 93.27 (PLs) 81.89, 80.97, 76.74 (NPLs)

Data-driven analytics is a technology based on data mining and artificial intelligence. The fundamental concept of this method is to utilize algorithms to make better decisions and predictions using the features extracted from the data. In addition, although a large amount of data and many variables are involved, data-driven analytics can be applied to complex tasks without pre-existing equations. Using these methods, studies are being conducted based on data from existing wells, and the key factors among various attributes are identified for future production behavior and productivity estimation.

Li and Han [9] performed principal component analysis (PCA), which is widely used for dimensionality reduction by minimizing data loss, for 10 factors, including reservoir properties and hydraulic fracturing design factors based on reservoir simulations; subsequently, seven key factors were utilized as input factors for the ANN models. Three variables of LGM, which is a DCA, were located in the output layer, and the ANN model was trained with 70% training, 15% verification, and 15% testing data. In addition, its applicability to the actual field site was determined by the mean square error (MSE) and correlation coefficient of 0.013 Mscf/day and 0.92, respectively. However, if there are no reservoir properties, it is difficult to apply the model; moreover, verification using field data is required.

He [10] identified the key performance indicator (KPI) using IMPROVE^{TM}, which was developed by Intelligent Solution Inc., for the reservoir characteristics, hydraulic fracturing design factors, and dynamic parameters, and developed an ANN model based on KPI. For 20 input layers, 80% training, 10% verification, and 10% testing data were learned using an error back-propagation algorithm, estimating 50 years of EUR, and analyzing the effect on the reservoir characteristics and hydraulic fracturing design factors on productivity. There are many input factors for the predictive model, and the future production volume can be obtained. However, this trend could not be identified.

Vyas et al. [11] considered four DCA methods using production rate data; three machine learning algorithms were applied to obtain the decline curve factors and EUR using nine factors by employing the productivity of shale formations as inputs. The results of averaging using Bayesian model averaging (BMA) and generalized likelihood

uncertainty estimation (GLUE) for various DCA results showed that the support vector machine (SVM) results of SEPD were the most accurate, followed by relative influence (RI) to input factors by initial production rate, total amount of proppant, and true vertical depth. Although several machine learning algorithms have been used, the accuracy of the prediction results should be improved.

Bowie [12] quantified multicollinearity by calculating the variance inflation factor (VIF) from correlation coefficients for factors related to maturity, pad wells, and hydraulic fracture size, and applied simple and multi-linear regression. By applying these as well as ANN, the prediction accuracy of the ANN model was the highest, with a coefficient of determination of 0.93. Although an individual well exhibits satisfactory performance, there are many input factors required to use the ANN model, which is difficult to apply.

Luo et al. [42] identified interrelated variables from correlation coefficients for the reservoir properties and hydraulic fracturing design factors and learned a deep learning model consisting of multiple hidden layers and neurons with eight input factors obtained from feature selection. However, despite using data from many production wells and models with many hidden layers and neurons, the coefficient of determination was 0.61. Furthermore, the effects of geological and hydraulic fracturing design factors on the production per unit length were identified using the developed prediction model. At low porosity, the thickness of the shale formation was high, the number of stages was large, and the production increased as the depth increased. Although many production wells were analyzed using deep learning, the results were unsatisfactory because data with different characteristics were not classified.

These studies show that using an ANN enhances prediction performance, and prediction using machine learning is possible in numerical analysis. Moreover, the selection and extraction process of the features that have a major impact on the prediction results should be performed. It must also be noted that some shale formations have high production variability, regardless of the number of production wells.

Unlike in conventional petroleum resources, existing analysis techniques have limited application for shale formations with uncertain flow mechanisms in production; therefore, the impact of hydraulic fracturing design factors on productivity must be identified. In addition, production forecasting for new wells is required, along with a development plan for well production. Recently, owing to the influence of oil and gas prices, drilled but uncompleted wells (DUCs) have been gradually increasing in North America [43], and their locations and depths have been identified.

Therefore, when developing new wells without information from shale reservoirs, the development plan is generally based on information from nearby production wells. Related development plans or productivity forecasting studies are in the early stages, and similar input factors from the existing production wells have been utilized for new wells to predict production behavior and cumulative production at a specific point in time.

Cao et al. [2] applied an ANN model to predict the production behavior of a single well, and tubing head pressure (THP) and production rate data were used to forecast the production of existing wells. For a new well, the results were obtained by inputting the target location, geological maps, THP, and production rate data obtained from nearby production wells; the results were compared with those obtained by applying the DCA of Arps, SEPD, Duong, and power law to more than four years of production rate data. Although the production behavior from DCA was similar to the actual data, the ANN model using THP was more accurate in the production forecast.

Mohaghegh et al. [15] introduced the concept of shale analytics, which can be applied to data-driven shale formations, using IMPROVE^{TM} to perform well quality analysis (WQA) and identify the relationship between productivity and native and design factors such as KPIs. This was estimated as an input factor of the ANN model for six months of cumulative production, with 96% accuracy for the existing production wells. For the prediction of new wells, the existing production wells were assumed to be blind, resulting in accuracy and average error rates of 77% and 12.9%, respectively. Furthermore, a look-back

analysis was performed using Monte Carlo simulations and prediction models to understand the influence of hydraulic fracturing design factors on production, and a method to present optimized hydraulic fracturing design factors for a single well was proposed.

Amr et al. [16] utilized a large amount of basin and other related data to estimate productivity for producing locations (PLs) and non-producing locations (NPLs). Using the Arps equation, which is an industry criterion, the initial decline rate and EUR were estimated, and the type curves of P10, P50, and P90 were obtained using information from nearby production wells. In addition, a machine learning model capable of estimating the initial decline rate and EUR was developed using data from nearby production wells that satisfy the conditions of the reservoir characteristics and hydraulic fracturing design factors of less than 2500 ft for production forecasting. Seven algorithms were applied, of which the EUR prediction accuracy for extreme gradient boosting trees (xgbTree) was 93.27% for PLs and 76.74% for NPLs. To improve the accuracy of NPLs, data from various regions were added and analyzed, but the highest case was 67.43%, which is relatively low in accuracy for non-production areas despite the use of reservoir characteristics and nearby production well data.

Therefore, in this study, the existing production wells were classified into groups according to the field conditions of shale formations, where various production characteristics were presented. The groups were then used to make reliable estimations of production. For new wells, there is a limit to the use of information from the nearby production wells in shale formations with severe vertical and horizontal changes; therefore, we utilize probabilistic inputs from different groups of production characteristics and perform output prediction considering the uncertainty in the new wells with limited data.

3. Data-Driven Analytics of Production Characteristics in Shale Formations

3.1. Feature Selection and Extraction for Identifying Key Factors

Productivity among different wells can differ based on geological properties or hydraulic fracturing designs [44], and limitations exist in the acquisition of reservoir properties and geological information among native, design, and dynamic parameters. Therefore, the features should be selected and extracted from data preprocessing from limited data, and related variables should be identified to derive key factors in shale formations.

It is advantageous to include sufficient and accurate information because machine learning depends on the quantity and quality of the data. However, the extent to which the accurate information can only be obtained through observation is limited. Nevertheless, this information can be obtained if the problem to be solved has a sufficient background of accurate information, which is not the case in most situations. Therefore, before applying a machine learning algorithm, it is necessary to select features that significantly influence the production performance. Feature extraction and selection were performed to check whether the features using sufficient data were valid. This method generates new input data based on existing inputs, and it is usually executed before the learning process, which is a key preprocessing step in the machine learning structure.

High-dimensional datasets have many correlating features and attributes. If these types of data are analyzed without characterization, problems such as overfitting and wasteful storage can occur. PCA is widely used in machine learning and pattern recognition techniques to solve these problems. Features can be extracted and selected by reducing the complexity and independence of the input parameters. PCA is a statistical procedure for generating relational variables for linearly uncorrelated vectors [9], and dimensionality reduction is possible with minimal data loss by analyzing multivariate data obtained for several response variables based on singular-value decomposition. This analysis technique has been applied to various fields for purposes such as history matching and seismic interpretation [45]. In addition, PCA provides new dimensions that can be used as the input variables for cluster analysis for efficient interpretation [46,47].

3.2. Unsupervised Learning for Group Classification

Unsupervised learning can be useful for unclear information contained in the data or for nonspecific objectives of the data search. Most unsupervised learning methods define groups based on measures of similarity or attributes for multiple objects with multiple attributes; this method is known as cluster analysis. This can be divided into hierarchical clustering and nonhierarchical or partitional clustering: hierarchical clustering has partial clusters within a cluster, whereas nonhierarchical clustering is mutually exclusive, without subsets or overlapping among clusters [48]. In addition, nonhierarchical clustering can be divided into hard clustering, in which each data point belongs to one category, and soft clustering, in which each measurement point can belong to two or more categories [49].

A typical hard clustering algorithm was used to determine the number of k clusters in advance. It sets the representative value of each cluster and assigns each object to one of the clusters based on the distance between the data and the centroid of each cluster. These methods are categorized based on representative values such as the mean or median. Soft clustering methods include fuzzy c-means and Gaussian mixture models, which are expressed as the probability that a single entity belongs to several clusters. Fuzzy clustering is similar to k-means cluster analysis but includes fuzziness, meaning that a point can belong to more than one cluster. The Gaussian model shows the probability that a point belongs to a cluster such that it can be easily used when the correlation structure of the cluster size and attributes varies [50]. Among these clustering methods, k-means can be used to determine the minimum linear distance from the center value, that is, to minimize the objective function. Each well is assigned to a unique cluster, which can lead to incorrect or biased results. The fuzzy c-means method represents the likelihood of an individual well belonging to several clusters, and the condition is obtained by updating the membership function repeatedly to determine a value that converges to the local minimum or saddle point of the objective function [51].

3.3. Supervised Learning for Classified Group Estimation

Supervised learning can train models with a set of labeled input and output data to make reasonable predictions of new inputs and applies classification and regression algorithms for model development. Common classification algorithms include decision trees, discriminant analysis, logistic regression, naïve Bayes, SVM, and k-nearest neighbor (k-NN) algorithms, which include linear and non-linear regression, SVM regression, and Gaussian process regression (GPR). Determining suitable methods among various algorithms requires sufficient data analysis and prior knowledge of the field.

Discriminant analysis identifies primary defects in features based on Gaussian distributions, classifies the data, and applies them to new data classification to generate class labels. Hereafter, straight lines or curves were created to classify the data into categories. Furthermore, discriminant analysis assumes that these variables are normally distributed and have the same covariance matrix in groups, and it is usually the first method to be utilized in classifier development [52].

The naïve Bayes classifier is a classification algorithm based on the concept of probabilities and statistics and is widely used in research [53]. This analysis is performed under the assumption that different features in clusters are statistically independent, based on conditional probability. Clusters are classified based on the highest probability that new data will belong to a specific cluster, and because it can be classified into training data, it has the advantage of high computational efficiency.

SVM classifiers identify a hyperplane that separates all measurement points from those of other clusters and classifies the data; the accuracy of the results depends on the maximum margin, which is the zone around the hyperplane where the distance is maximum in each nearest plane. In nonlinear cases, it is applied by converting into higher dimensions, where hyperplanes can be found using kernel transformations [54]. SVM regression algorithms function similarly to SVM classification algorithms but are modified to predict continuous responses, which identify models that deviate by small values from

the measured data using as few parameters as possible to minimize sensitivity to errors instead of exploring the hyperplane.

k-NN categorizes objects based on nearby clusters within the dataset, and distances, such as Euclidean and cosine, are used considering the assumption that objects close to each other are similar. It is the simplest non-parametric procedure to determine labels for data samples, and the nearest k neighboring labels are determined based on the voting mechanism [55].

Among the regression algorithms, linear regression expresses continuous response variables as a linear function for one or more predictors; although the model is simple, it can provide an appropriate and interpretable description of the relationship between inputs and outputs [56]. This method is universally utilized to obtain the least-squares method.

3.4. Artificial Neural Networks for Production Forecasting

ANN is expressed as a mathematical connection relationship between nerve cells by mimicking the operating principle of the nervous system and biological neurons and comprises an input layer, a hidden layer, and an output layer. It is learned by varying the intensity of the connection. The output of the neuron can be changed according to the type of activation function, as follows:

$$a^i = \sigma\left(W^i a^{i-1} + b^i\right) \tag{2}$$

where a^i is the neuron in layer i, a^{i-1} is the neuron in layer $i-1$, b^i is the bias vector, W^i is the weight matrix for each layer i, and σ is the activation function.

Activation functions are typically binary, linear, rectified linear, sigmoid, Gaussian, or hyperbolic tangent functions. The binary function is a unipolar or bipolar function; when the sum of input weights is less than the threshold, the output of the neuron is 0, and the output of the neuron is 1 when it is greater than or equal to the threshold. The sigmoid function is a unipolar or bipolar nonlinear continuous function, and any form of input value can be expressed as a value between 0 and 1.

For data, an ANN can apply supervised or unsupervised learning. In the former case, the connection strength is changed such that the output value is within the error range of the target according to the given input, and typical models include Perceptron, Hopfield, and backpropagation. In the latter case, similar input patterns are adjusted to be learned at the same output, including Kohonen's competitive learning and Grossberg's adaptive response theory (ART) model.

ANNs can derive pattern recognition and nonlinear features, despite the incomplete learning process of the input data, and can process a large amount of data simultaneously in a short time [57]. Furthermore, by adjusting the connection strength through learning, the relationship between the input and output data can be identified without prior knowledge. Therefore, we can model nonlinear systems, even if the model is to be continuously updated.

4. Analysis of the Existing Wells

4.1. Acquisition of Field Data and Preprocessing

The field data were normalized with the lateral length or number of stages to allow comparisons among multiple wells (Figure 4). The normalized volume (NV) and normalized production (NP) were derived, and the value per unit length was calculated assuming that the number of stages was equally distributed [42]. In addition, preprocessing was performed on production-related attributes, such as removing very small or large outliers for each attribute.

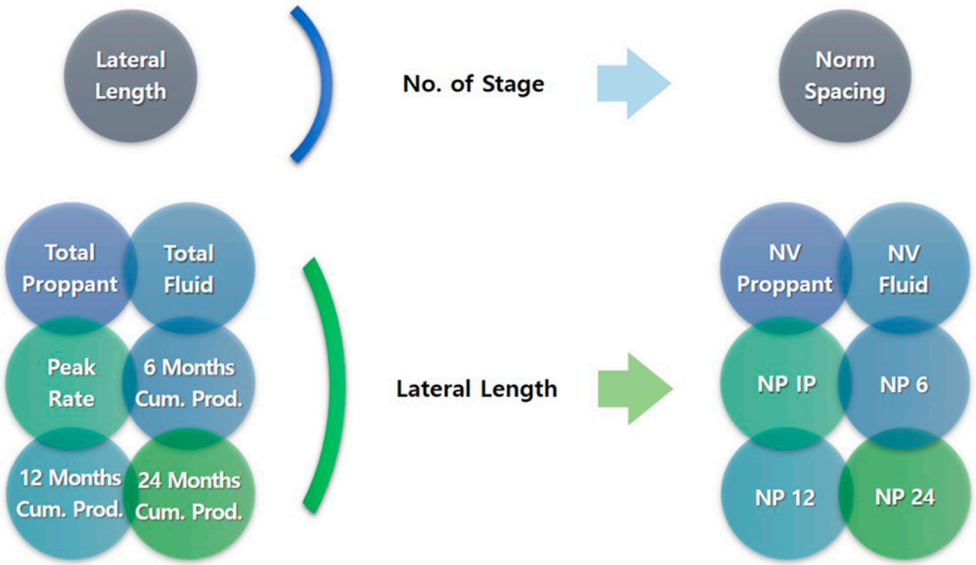
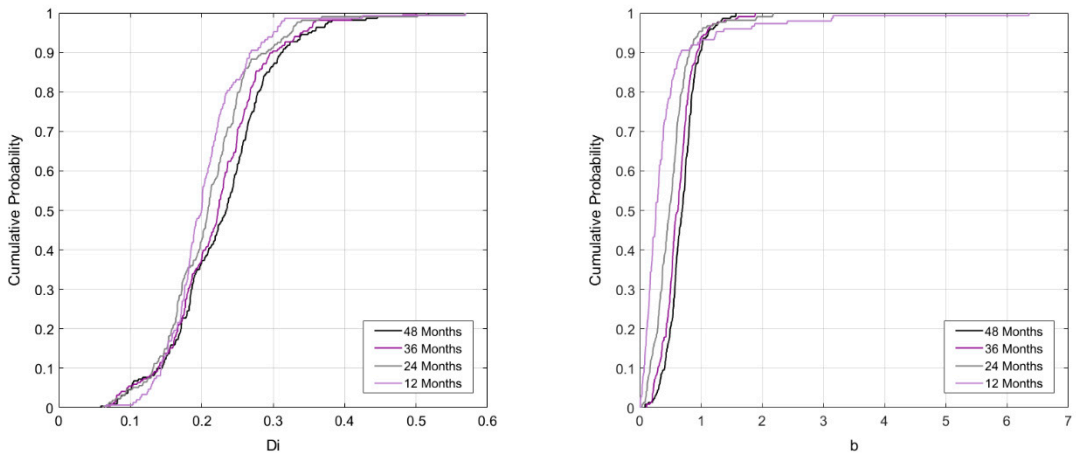


Figure 4. Data preprocessing for shale formations.

Normalization was performed to obtain NV and NP for comparison of multiple wells, and preprocessing was performed on production-related attributes in the study area. To obtain the dynamic attributes, previous studies used more than 12 months to estimate the decline curve parameters [16,19], but stable production behavior was observed when long-term production data were used [33]. In this study, 48 months of production history data from 220 production wells were used to obtain the dynamic attributes. Different results of the decline exponent were estimated according to the period of production rate data used for DCA [35]. Therefore, DCA was performed according to the production period, and the cumulative probability distribution of the initial decline rate and decline exponent was derived for the total production wells (Figure 5). The production rate data for a sufficient period of 48 months were used; however, despite the shale formations, a value exceeding 1 rarely appeared, (Figure 6). If the decline exponent exceeds 1, it is called 'beyond hyperbolic', which is in the same form as that of the hyperbolic function in a shale reservoir [37].



(**a**) (**b**)

Figure 5. Change of cumulative probability for the DCA parameters according to production period: (**a**) the initial decline rate; (**b**) the decline exponent analysis for permeability.

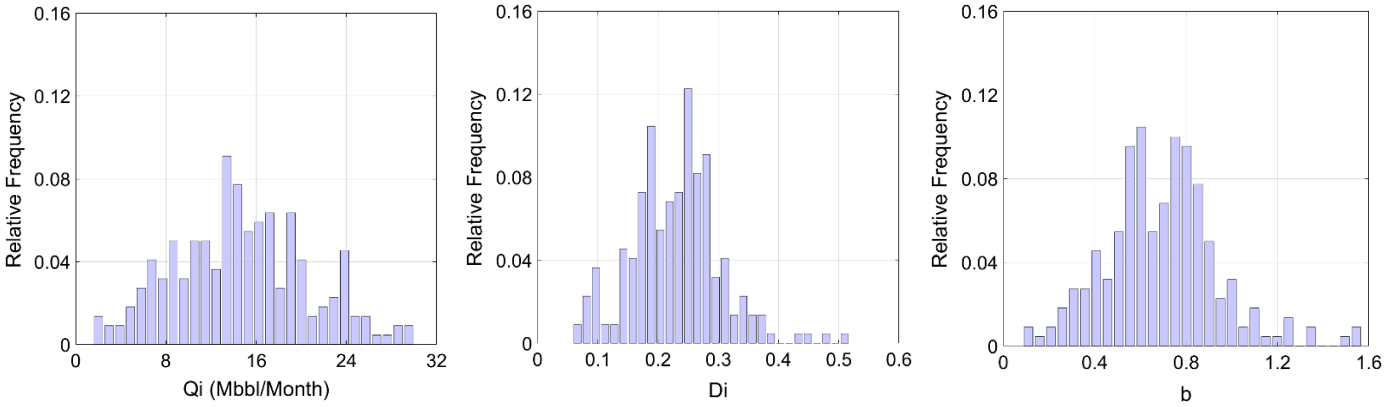


Figure 6. Relative frequency of the DCA parameters after preprocessing: (**a**) the initial production rate; (**b**) the initial decline rate; and (**c**) the decline exponent.

4.2. Identification of Correlations and the Key Factors

The 22 attribute datasets obtained in this study based on the preprocessing and DCA results are summarized in Table 3. In the case of the landing direction, the azimuth angle was derived from the position of the production well at the surface and bottom hole, which can be referred to when generating hydraulic fracturing cracks in the direction of the horizontal well. Correlations were analyzed to identify the relationships among production-related attributes. For the four attributes of the well location and five attributes associated with production, as in Figure 7, similar factors indicate a high correlation. Therefore, the correlation coefficients are shown in Figure 8 for factors that show trends, except for attributes that are not correlated and are similar.

The lateral length and measure depth (MD) showed a correlation of 0.8, and the hydraulic fracturing (HF) stage, lateral length, norm spacing, and NV fluid (NVF) showed a correlation exceeding 0.7. In addition, the MD and HF stages, NV proppant (NVP), norm spacing, and NVF and NVP showed values over 0.6. Finally, the HF stage and norm spacing exhibited correlations of >0.5. However, the other factors were found to have nonlinear relationships.

Surface Latitude	0.9802	0.9982	0.9805
0.9802	Surface Longitude	0.9814	0.9995
0.9982	0.9814	Bottomhole Latitude	0.9807
0.9805	0.9995	0.9807	Bottomhole Longitude

(**a**)

NP IP (bbl/ft)	0.9332	0.8876	0.8462	0.9873
0.9332	NP 6 (bbl/ft)	0.9835	0.9505	0.9766
0.8876	0.9835	NP 12 (bbl/ft)	0.9870	0.9429
0.8462	0.9505	0.9870	NP 24 (bbl/ft)	0.9025
0.9873	0.9766	0.9429	0.9025	Arps NP Qi (bbl/ft)

(**b**)

Figure 7. Highly correlated input attributes: (**a**) the attributes of the well location; (**b**) the attributes of production.

Figure 8. Correlation coefficients among input attributes.

Table 3. List of acquired attributes in this study.

Native	Design	Dynamic
Surface latitude	Lateral length (ft)	NP IP (bbl/ft)
Surface longitude	HF stage	NP 6 months (bbl/ft)
Bottom hole latitude	Norm spacing (ft/stage)	NP 12 months (bbl/ft)
Bottom hole longitude	NVP (lbs/ft)	NP 24 months (bbl/ft)
Measured depth (ft)	NVF (bbl/ft)	NP Q_i (Arps) (bbl/ft)
True vertical depth (ft)		D_i (Arps)
Landing direction (degree)		b (Arps)
Choke size		
County number		
Well elevation (ft)		

Well productivity depends on the native parameters of shale formations and the success of the hydraulic fracturing design. PCA was conducted to reduce and summarize the correlated multidimensional variables and identify the key factors among the limited datasets. Factors that have a major influence on the various attributes related to production were identified as key factors in this study. Four case studies and PCAs were performed for the input attributes to identify the key factors, as listed in Table 4.

Among the 22 attributes, if the correlation between the production volume and well location exceeded 0.9, it was considered a duplicate state. Duplicate values were eliminated from the 22 attributes, and 16 attributes were analyzed for Case 1, considering the surface latitude, surface longitude, and NP 6. For Case 1, it was found that three principal components (PCs) represented approximately 51% of the total attributes, and the PC coefficients of county number, well elevation, landing direction, and choke size were smaller than those of the other attributes (Figure 9a). In Case 2, 12 attributes were analyzed, except for the four attributes in Case 1, which had a low correlation with the PCs. Three PCs represented approximately 76% of the total attributes (Figure 9b). In Case 3, two decline curve parameters with relatively small PC coefficients and well locations with similar ranges of coefficient values for all PCs were excluded. The results revealed that the three PCs represented approximately 82% of the eight attributes, and the effects on the main attributes were considered for Cases 1 and 2 (Figure 9c).

Table 4. List of input attributes in the case studies.

Case 1	Case 2	Case 3	Case 4
County num.	Surface latitude	MD	TVD
Surface latitude	Surface longitude	TVD	HF stage
Surface longitude	MD	Lateral length	Norm spacing
Well elevation	TVD	HF stage	NVP
Landing direction	Lateral length	Norm spacing	NVF
MD	HF stage	NVP	
TVD	Norm spacing	NVF	
Lateral length	NVP	NP 6	
HF stage	NVF		
Norm spacing	NP 6		
NVP	Di (Arps)		
NVF	b (Arps)		
NP 6			
Choke size			
D_i (Arps)			

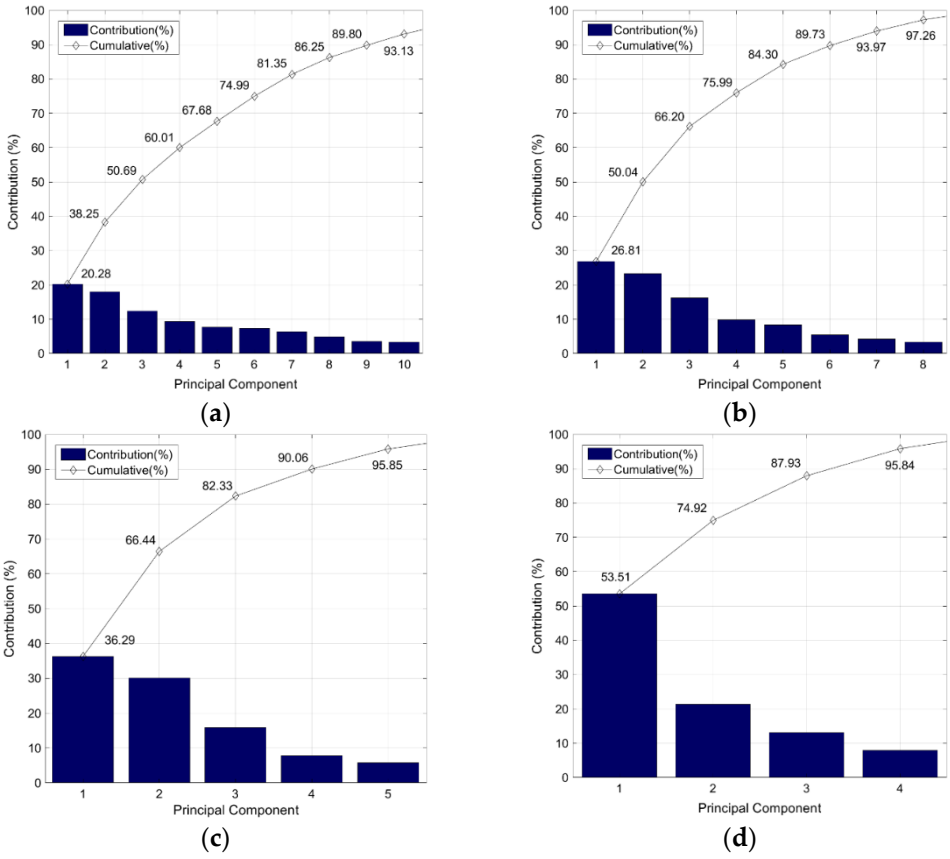


Figure 9. Scree plot of PCA results for all cases: (**a**) Case 1; (**b**) Case 2; (**c**) Case 3; (**d**) Case 4.

Apart from the lateral length used for normalization, based on the analysis of the primary attributes affecting the PCs, a PCA was performed in Case 4 to determine the effects of the true vertical depth (TVD) and hydraulic fracturing design factors. As shown in Figure 9d, it was found that the three PCs represented a large portion (approximately 88%) when using only five limited attributes. In addition, Table 4 shows the values of the key factors for each PC, and based on these values, factors with absolute values of 0.4 or higher were determined to be correlated with the PC. Moreover, the influence of the key factors on each PC was quantitatively different and classified without overlapping. These results are shown in Figure 10, and a significant correlation between the positive

and negative directions was observed. The norm spacing, NVF, and NVP for PC 1, TVD and HF stages for PC 2, and the number of HF stages and NVP for PC 3 are represented as follows:

$$PC1 = 0.1146\ TVD + 0.3921\ Stage - 0.5601\ Spacing + 0.4803\ NVP + 0.5374\ NVF \quad (3)$$

$$PC2 = 0.8905\ TVD - 0.4446\ Stage - 0.0111\ Spacing + 0.0396\ NVP + 0.0876\ NVF \quad (4)$$

$$PC3 = 0.3957\ TVD + 0.7112\ Stage - 0.0583\ Spacing - 0.5525\ NVP - 0.1702\ NVF \quad (5)$$

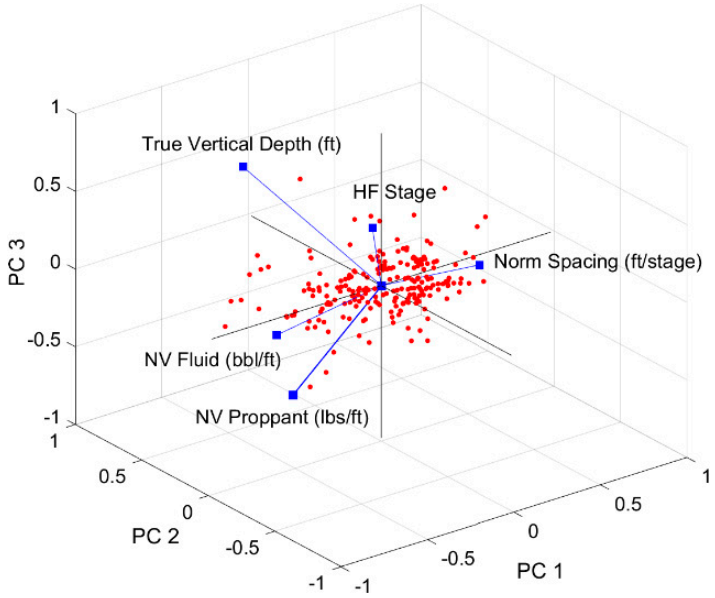


Figure 10. Visualization of the PCs in 3D plot for Case 4.

The case study found that the key factors among the available data in the study area were TVD and hydraulic fracturing design factors. In addition, depending on the study area or acquired data, the key factors may be different, according to the PCA results, which could be used for group classification based on the identified key factors.

4.3. Analysis of Production Characteristics Using Group Classification

To classify the groups based on the PCA results, fuzzy c-means were used, which indicates the possibility that an individual belongs to several clusters and is considered one of the most representative fuzzy cluster analyses. In this study, fuzzy cluster analyses were performed 10 times using the three PCs of Case 4, which consisted of the key factors from the PCA. This was verified by confirming that the objective function converged to a constant value in all trials. The number of wells in the three groups was 50, 74, and 96, respectively (Figure 11). This allows each group to identify a distinct classification of hydraulic fracturing design factors by PC 1, which is correlated with the norm spacing, NVF, and NVP.

To compare the production among wells, the average and cumulative production by group were determined based on the normalized production per unit length (Figure 12). It was found that the initial and cumulative production decreased in the order of Clusters 1, 2, and 3. In addition, the decline trend for up to approximately 15 months differed from group to group, whereas the trend in the latter half of production was similar. In particular, the production of Cluster 3 was small, and it would be better to exclude cases belonging to Cluster 3 when developing new wells in the study areas.

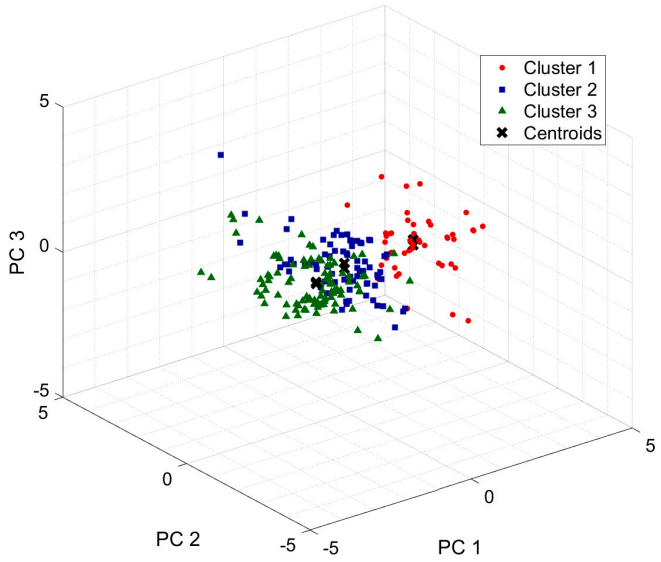
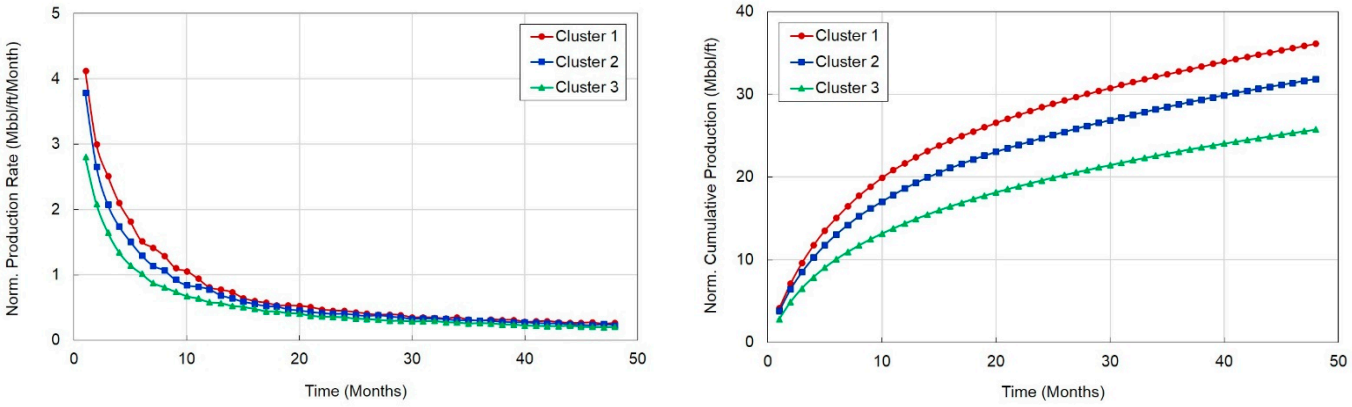


Figure 11. Cluster analysis results.



(**a**) (**b**)

Figure 12. Normalized production of each group: (**a**) the average production rate; (**b**) the average cumulative production.

To determine the group features, a statistical analysis of the key factors was performed to draw box plots of the mean, maximum, median, quartile, and minimum values for each key factor (Figure 13). First, we analyzed the results for each key factor. TVD was similar to Clusters 1 and 2, and Cluster 3 was in a shallower area than Clusters 1 and 2. Compared to the normalized production results, the deeper the TVD in the study area, the better the productivity. The HF stage had many stages of Cluster 1, and Clusters 2 and 3 were similar. The higher the number of HF stages, the better the production; however, it is important to determine the appropriate number of stages. Norm spacing is more productive if narrow, but it is necessary to identify appropriate intervals where interference does not occur. If only the norm spacing is considered, Cluster 1 was the narrowest, thus leading to a large production. It is known that productivity increases with increasing amounts of NVP and NVF, and Cluster 1 had the largest amount; therefore, its production is expected to be large. NP 6 is the cumulative production per unit length at six months and is plotted to check the difference in the early stages of production, independent of the key factors. In the case of normalized production, Clusters 1 and 2 had a slight difference in production, and Cluster 3 had the least.

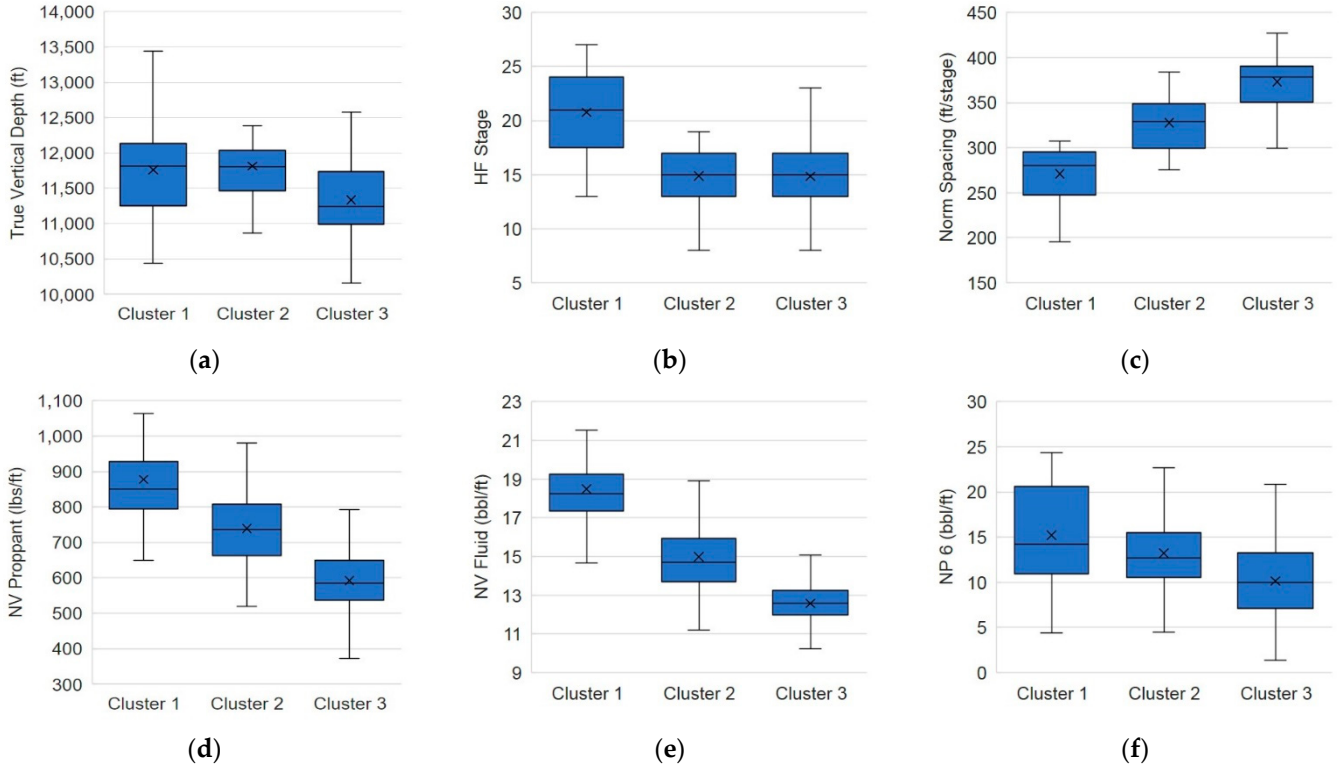


Figure 13. Variable distribution of the three groups based on key factors: (**a**) the TVD; (**b**) the HF stage; (**c**) the norm spacing; (**d**) the NVP; (**e**) the NVF; (**f**) the NP6.

These results revealed that hydraulic fracturing increased the production volume for narrow spacing and large amounts of injected proppant and fluids per unit length. Although Clusters 1 and 2 had similar TVD, they contained wells with higher productivity under the influence of hydraulic fracturing design factors. Clusters 2 and 3 had similar numbers of HF stages, but the cumulative production per unit length was high owing to the hydraulic fracturing design factors. Had the hydraulic fracturing design factors increased, the productivity of Cluster 1 would have been good, but there were production wells with similar productivity in Cluster 2. This means that Cluster 2 performed hydraulic failure under worse conditions than Cluster 1, but Cluster 2 was more productive on similar TVD.

Unlike the previous studies, in which production-related attributes were not discriminated by production volume [7,11], the results obtained by categorizing the groups using the key factors of shale formations showed different ranges of hydraulic fracturing design factors. This enabled us to verify the results of the key factors derived using PCA and showed that the procedure proposed in this study is applicable to more than grouping only. Furthermore, it was possible to identify the field conditions of high productivity in the study area. The feature values are expected to be used for classification of groups with different production characteristics for new wells.

4.4. Production Estimation Using Classified Groups

If the initial production data for the shale formations are available, DCA is usually performed to obtain the production forecast and estimate the EUR or cumulative production for a certain period (i.e., 48 months or long-term period). Production was estimated for 48 months using the production data for 12 months. After the Arps hyperbolic equation was applied to the production rate data for 12 months, the distribution of the estimated NP 48 for cumulative production was obtained, as shown in Figure 14. Outliers were excluded, and the average absolute percentage error (AAPE) was calculated as follows:

$$AAPE = 100/n \sum_{i=1}^{n} |Measured_i - Estimated_i / Measured_i| \quad (6)$$

where n represents the number of datasets. The results revealed that the error was approximately 26% for all production wells, and relatively large errors were obtained owing to the production variability over a short production period. When six months of production rate data were applied to DCA, many production wells yielded abnormal estimates and were excluded from the comparison.

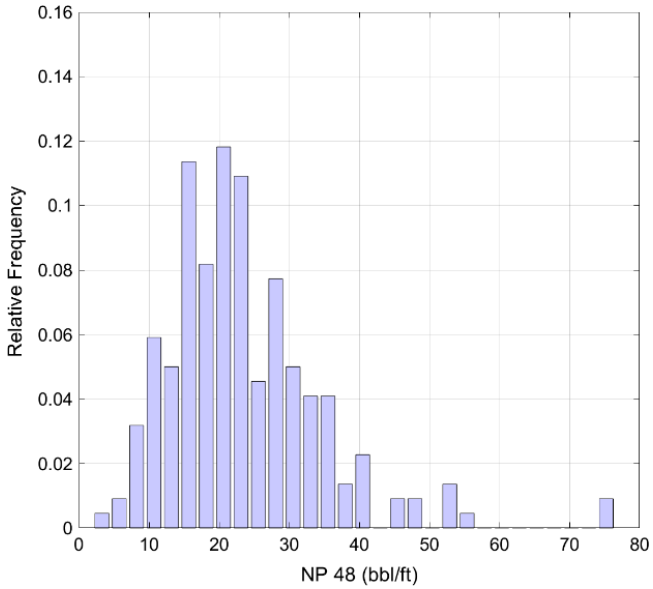


Figure 14. Estimated production results of all wells using the Arps hyperbolic for 12 months.

The normalized cumulative production for 48 months was estimated by deriving trend lines from the normalized cumulative production for 6 and 12 months for each group. Figures 15 and 16 show the regression equation and determination coefficient of each trend line, respectively. The AAPE of the estimated values using the trend line for each group is presented in Table 5. When the trend line is calculated with NP 6, the coefficient of determination (R^2) is lower than that of NP 12, but it has a value near 0.8; hence, it can be used to estimate production. The coefficient of determination for the trend line of NP 12 is 0.9; hence, it is recommended to use the trend line if the production data for the period can be utilized. In addition, the accuracy of the trend line can be evaluated using error analysis, such as the AAPE. When calculated with NP 6, it has a value of approximately 10%, and for NP 12, it shows an error of less than 8%. This is a much better and reliable result than the 26% error that was derived when DCA was applied to the production rate data for 12 months of the entire wells. Furthermore, it was difficult for DCA to analyze the production rate data for six months.

If there is a cumulative production for 6 or 12 months for a producing well, it is possible to predict the cumulative production for 48 months using the trend line derived from the input data characterization. The cumulative production at different times can also be predicted using the trend line. It can be seen from the graph that Cluster 3 could be used to predict production wells with relatively low productivity, and that the productivity of Clusters 1 and 2 was similar.

Table 5. AAPE of the estimated production results by group.

	Cluster 1	**Cluster 2**	**Cluster 3**
NP 6 vs. NP 48	9.72%	12.02%	12.60%
NP 12 vs. NP 48	6.44%	7.94%	7.87%

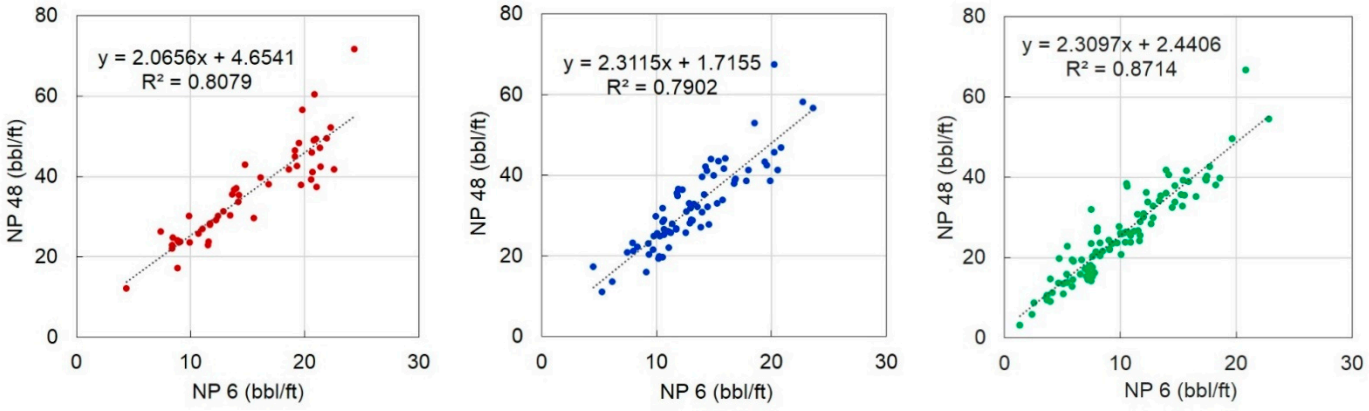


Figure 15. Relationship between NP 6 and NP 48 by group: (**a**) Cluster 1; (**b**) Cluster 2; (**c**) Cluster 3.

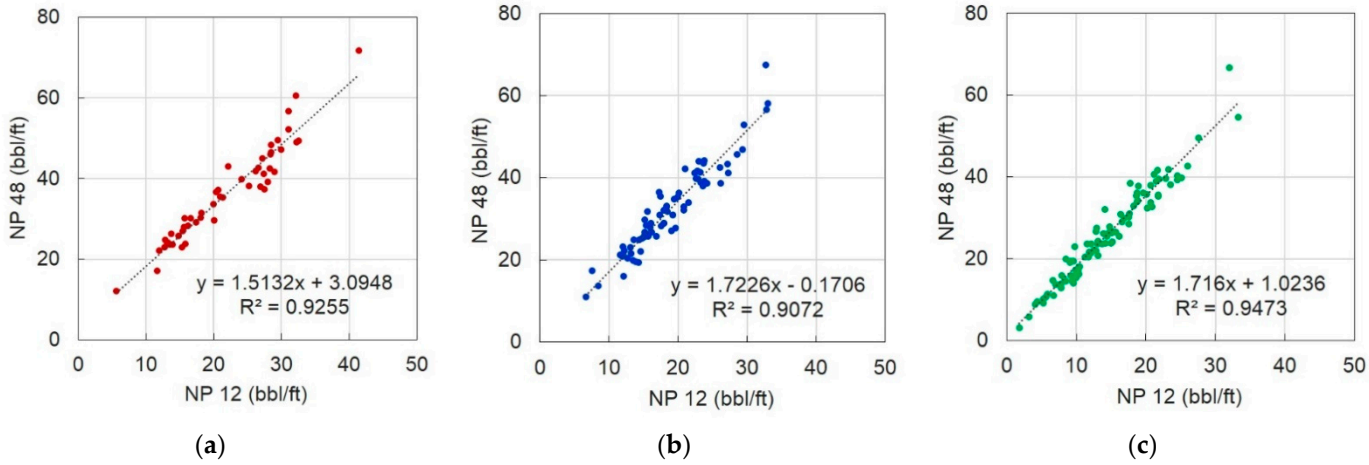


Figure 16. Relationship between NP 12 and NP 48 by group: (**a**) Cluster 1; (**b**) Cluster 2; (**c**) Cluster 3.

5. Analysis of the New Wells

This section aimed at applying various methods for predicting production from new multi-wells with limited input factors, utilizing databases from existing wells. Here, we classified groups with different production characteristics and predicted probabilistic results rather than single values for uncertainty consideration of the forecasts.

5.1. Probability Input and Feature Extraction by Group

We utilized the characteristics of classified groups for reliable production forecasting in new wells, and identified statistical features for the key and decline curve factors related to production among the 22 attributes. From the box plot, we can see that each key factor normalized between 0 and 1 for each group had different characteristics of the mean, maximum, median, minimum values, and interquartile range (IQR) (Figure 17). Among them, the group characteristics of spacing, NVP, and NVF were clearly distinguished among the hydraulic fracturing design factors and were expected to be available as a major attribute in group classification. Furthermore, due to differences with other groups on the characteristic values of the hydraulic fracturing design factors of Cluster 1, it can be identified as containing high-productivity wells at initial production. Assuming that the production well used in the analysis applied the optimal hydraulic fracturing design, the grouping would enable the estimation of hydraulic fracturing design factors for new wells.

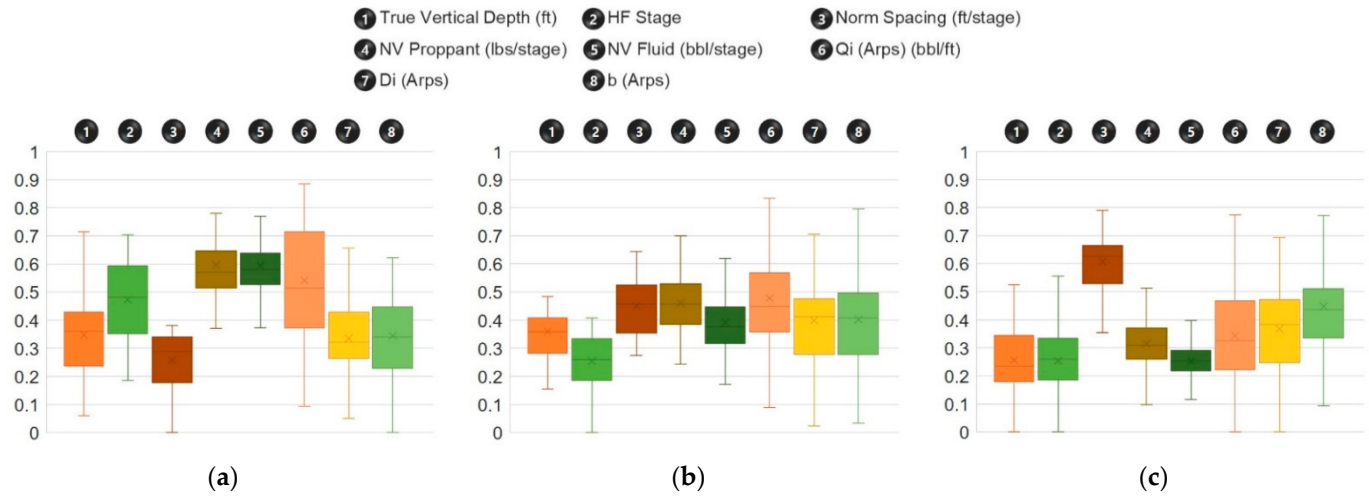


Figure 17. Normalized variable distribution of key and production factors by group: (**a**) Cluster 1; (**b**) Cluster 2; (**c**) Cluster 3.

Owing to the characteristics of the heterogeneous shale formations when selecting input values for new wells, uncertainties exist when predicting production only with information on nearby existing production wells. To take this into account, we derived probabilistic values (*p*-value) for each group to utilize the input parameters of the prediction model, P10, P50, and P90, summarizing them in Table 6 from the cumulative probability graph of Figure 18. From these results, the production characteristics classified by the influence of the key factors reflect the features of different groups and are expected to be available for predicting the production of new wells.

Table 6. Probability value of the DCA parameters by group.

	Cluster 1			Cluster 2			Cluster 3		
	P10	**P50**	**P90**	**P10**	**P50**	**P90**	**P10**	**P50**	**P90**
Initial production rate (bbl/ft/Month)	1.95	3.25	4.97	1.99	2.89	4.57	1.03	2.18	3.73
Initial decline rate	0.10	0.20	0.28	0.15	0.25	0.32	0.14	0.23	0.31
Decline exponent	0.25	0.59	0.88	0.33	0.68	0.99	0.49	0.73	1.02

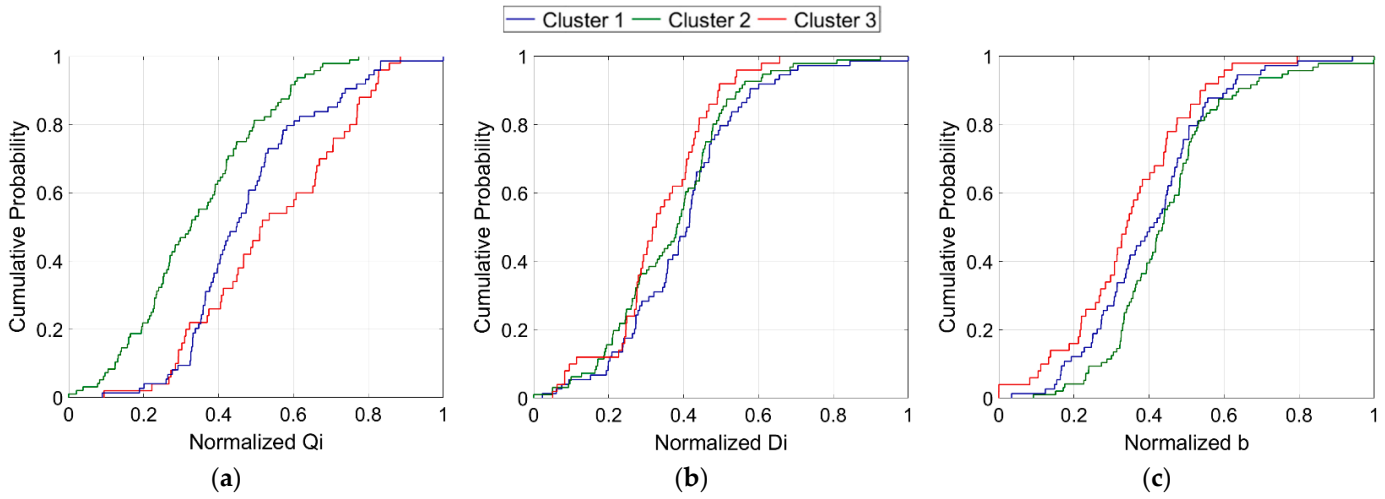


Figure 18. Cumulative probability of the DCA parameters by group: (**a**) Cluster 1; (**b**) Cluster 2; (**c**) Cluster 3.

5.2. Case Study of Input Attributes for Group Estimation

In this study, we developed a model using machine learning to classify groups using the key factors without going through a series of processes with different production characteristics of the shale formations. To this end, we applied the classification algorithms among the supervised learning methods to organize the input data into 80% for training and 20% for testing, and performed five-fold cross-validation on training.

Considering the well location, case studies were carried out according to the input variables (Table 7); the input variables were intended to utilize the attributes of the key factors and dynamic parameters. For the case of six and seven attributes, the dynamic attributes were excluded, and the case of four attributes was analyzed without considering TVD.

Table 7. Case study for input attributes.

8 Attributes	...	5 Attributes	...	3 Attributes
TVD		TVD		Norm Spacing
Lateral Length		HF Stage		NVP
HF Stage		Norm Spacing		NVF
Norm Spacing		NVP		
NVP		NVF		
NVF				
NP IP				
NP 6				

The results with the highest accuracy for each attribute count are presented in Figure 19 with different classification algorithm results owing to the effect of the attributes. Among the classification algorithms, the accuracy of discriminant analysis, naïve Bayes, and SVM was approximately 80%, and the results depended on the attributes used by each algorithm. In the case of discriminant analysis, the accuracy of the test data was high in six to seven attribute cases that did not contain dynamic attributes, and SVM showed high prediction accuracy in four attribute cases, with the exception of TVD. The naïve Bayes classifier did not have a significant change in accuracy due to its attributes, but among them, the prediction accuracy was high for three to five attribute cases that included the key factors. This allowed us to identify the different classification algorithms with high accuracy depending on the effect of the attributes, and the group classification results of the existing production wells in which all the information exists were relatively less affected by the input variables.

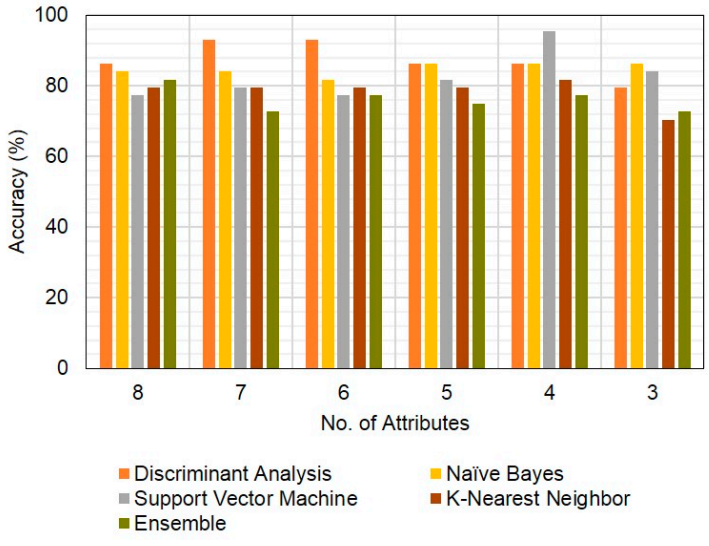


Figure 19. Accuracy for testing wells by machine learning methods and number or attributes.

The classification model verified with existing production wells was used to classify 44 blind wells (new wells). For the missing attributes in new wells other than well location and TVD, the average value of each attribute in the nearby production wells within a distance of 2500 ft was utilized, referring to the candidate conditions of Amr et al. [16].

We also performed case studies according to the input variables for the new wells, which resulted in similar algorithm accuracy, depending on the number of attributes, as shown in Figure 20. However, unlike the testing with existing production wells, classification by group was clearly indicated using three attributes: spacing, NVP, and NVF among the key factors, and the accuracy of the prediction result was the highest. This was considered to be the result of uncertainty due to the use of the average value of nearby production wells for input data on new wells. In addition, the error and accuracy results for each group of three attributes are shown in Figure 21; the accuracy of the naïve Bayes algorithm was approximately 70%. In the validation of classification models using existing production wells, naïve Bayes also had high accuracy according to the attributes. Moreover, this algorithm, which classifies groups based on probabilities, is appropriate when using limited input attributes.

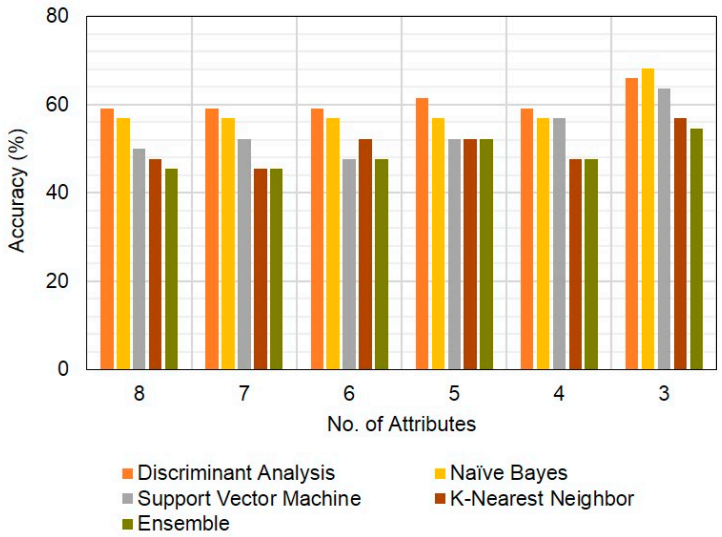


Figure 20. Accuracy for the new wells by machine learning methods and number of attributes.

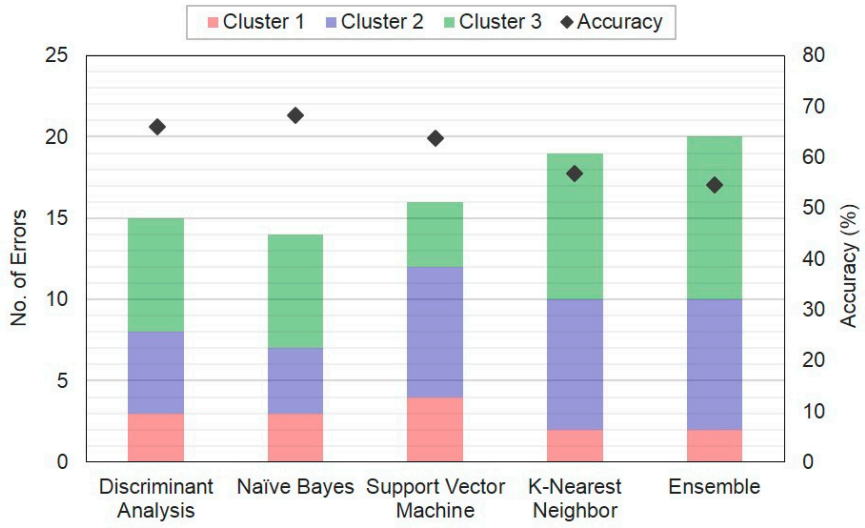


Figure 21. Classification results of group by machine learning methods for 3 attributes.

For the new wells, case studies were performed according to the input variables for group classification with different production characteristics, which allowed us to identify different algorithms with high accuracy, depending on the attributes and data features. Thus, performing a data-driven analysis may change the appropriate algorithm owing to the influence of the input variables or data features utilized in the analysis. Furthermore, we confirmed that group predictability is possible by using limited attributes and including key factors. The reliability improvement in production prediction can be achieved by classifying groups of new wells.

5.3. Development and Validation of ANN Model

The input layer of the ANN models, as shown in Figure 22, considered the new wells with limited information, with analysis results of existing production wells, utilizing the location and TVD of wells, spacing, and decline curve factors of the Arps equation. In the development of ANN models, we optimized them by considering the number of neurons in the hidden layers, performing a validation check, and epoch, and applying normalized input values between 0 and 1 for numerical stability.

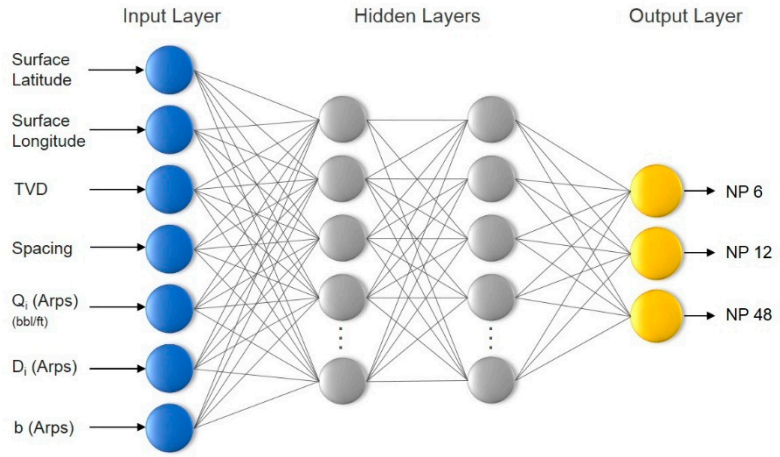
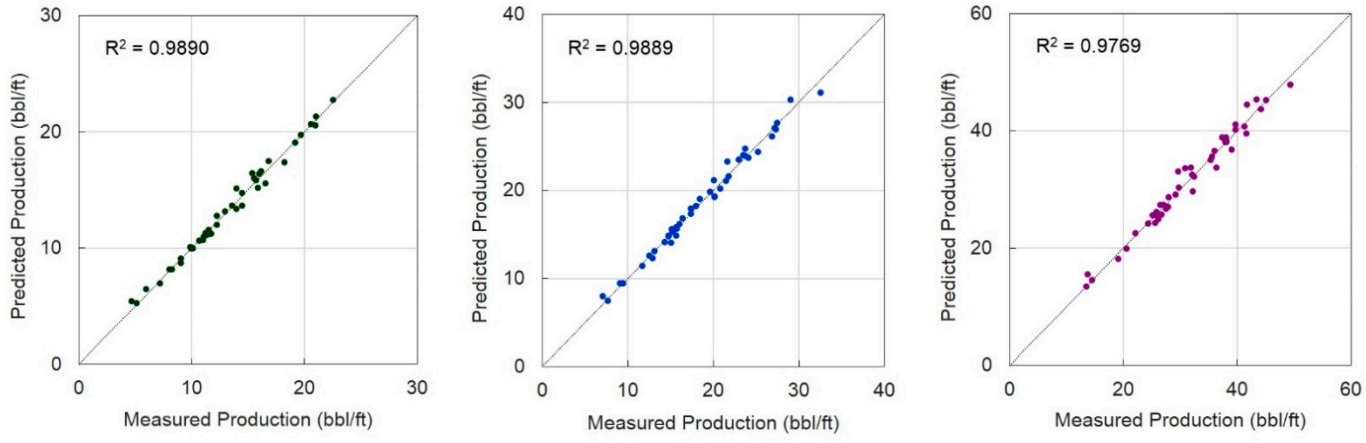


Figure 22. Structure of the ANN model in this study.

A model with 12 hidden layers and a learning function of scaled conjugate gradient (SCG) that can update the connection strength by back-propagating the error signal was determined. For each group, the available data were split into training (including validation) and test data with a ratio of 80(20):20. A bipolar sigmoid function was applied between the input layer and the first hidden layer between the first and second hidden layers, and a linear function was used between the second hidden layer and the output layer.

To verify the ANN model, we compared the estimated results of the existing production wells with the actual production (Figure 23). Consequently, NP 6, 12, and 48 for the 44 blind wells had AAPEs of 2.85%, 2.69%, and 3.24%, respectively, with approximately 98% accuracy. This enables the estimation of cumulative production per unit length for multiple wells; this can be used even in the case of oil wells that have been produced for a short period.

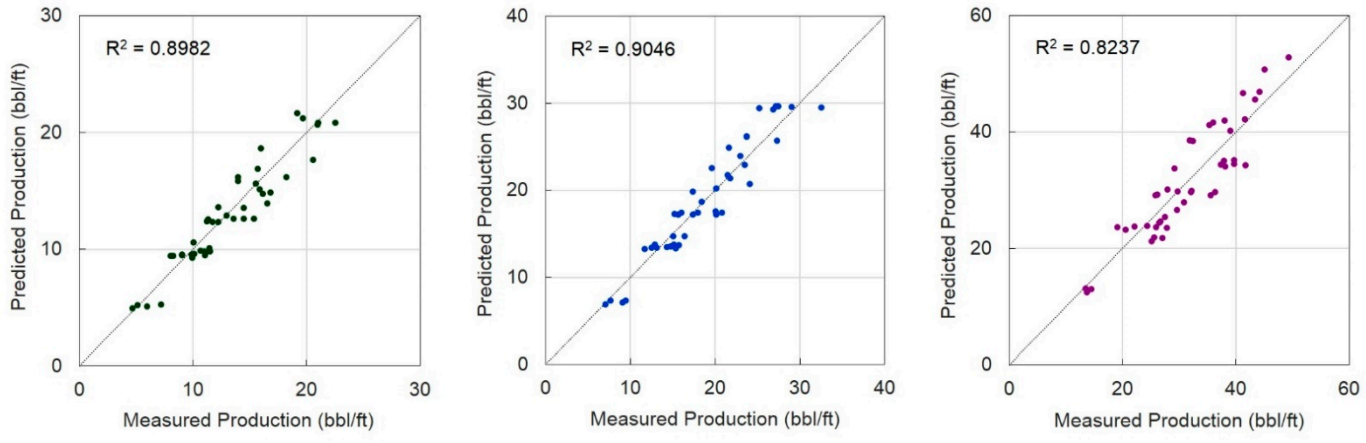


(**a**) (**b**) (**c**)

Figure 23. Comparison of the measured and predicted production for the testing wells: (**a**) NP 6; (**b**) NP 12; (**c**) NP 24.

5.4. Prediction of Probabilistic Production

In the ANN model validation, the well location, TVD, and decline curve factors used values that changed depending on the well. However, for new wells, production was predicted by utilizing three *p*-values for the spacing and decline curve factors to consider the uncertainty of input factors. For 44 blind wells, the AAPE with the minimum error of the three results of P10, P50, and P90 for each group were calculated as 9.52%, 8.85%, and 11.01%, respectively. Furthermore, we compared the actual production with the prediction results of the new wells using a validated ANN model (Figure 24) with an accuracy of approximately 80% or more.



(**a**) (**b**) (**c**)

Figure 24. Comparison of the measured and predicted production for the new wells: (**a**) NP 6; (**b**) NP 12; (**c**) NP 24.

The results of probabilistic production forecasting for each well are shown in Figure 25, and the cumulative production at 48 months showing relatively stable production behavior was included between the P10 and P90 results. In the 6- and 12-months cases, production variability was large in the shale formation, so the range of forecast results did not include some cases of large or small production. However, it is close to the range of the probabilistic production.

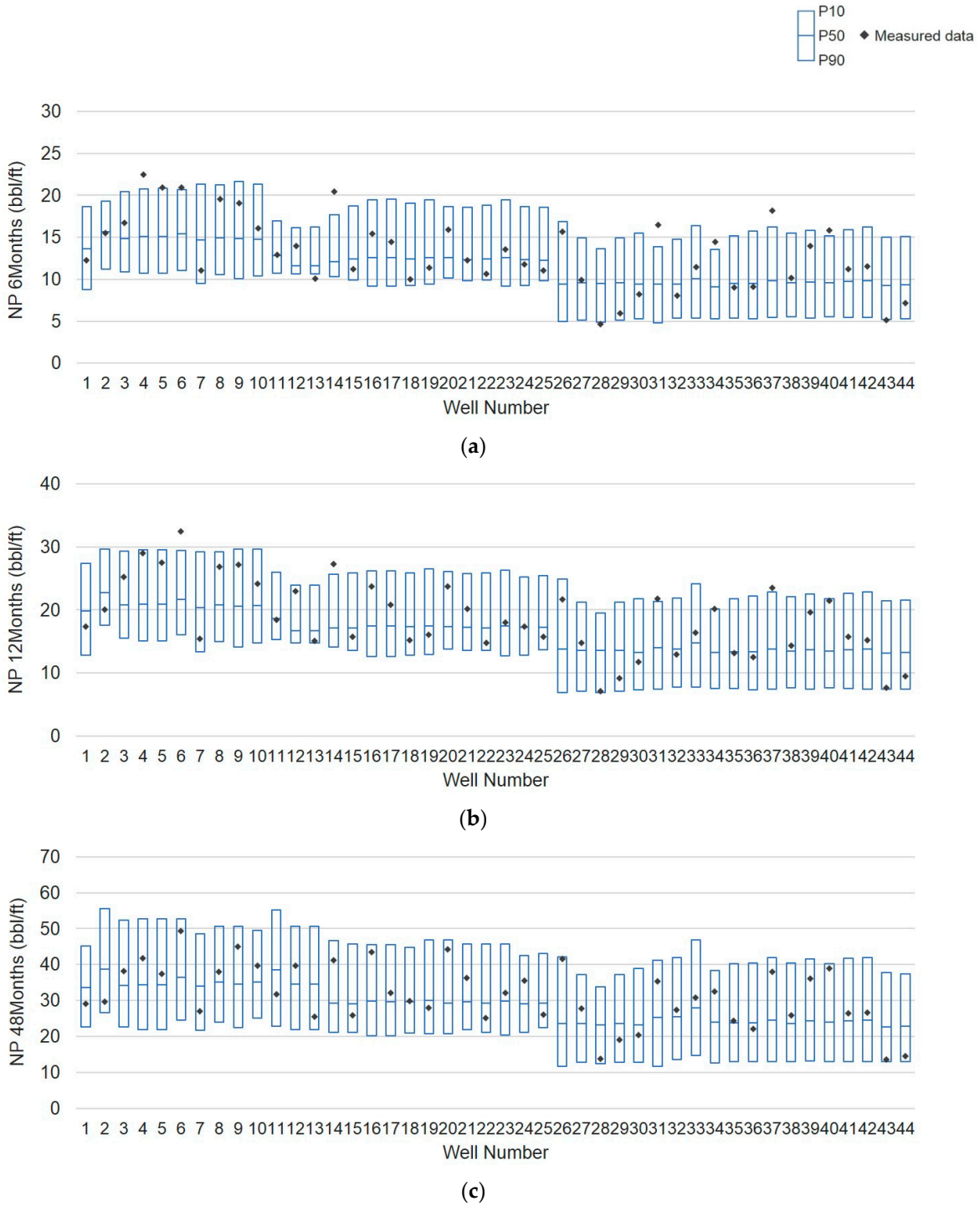
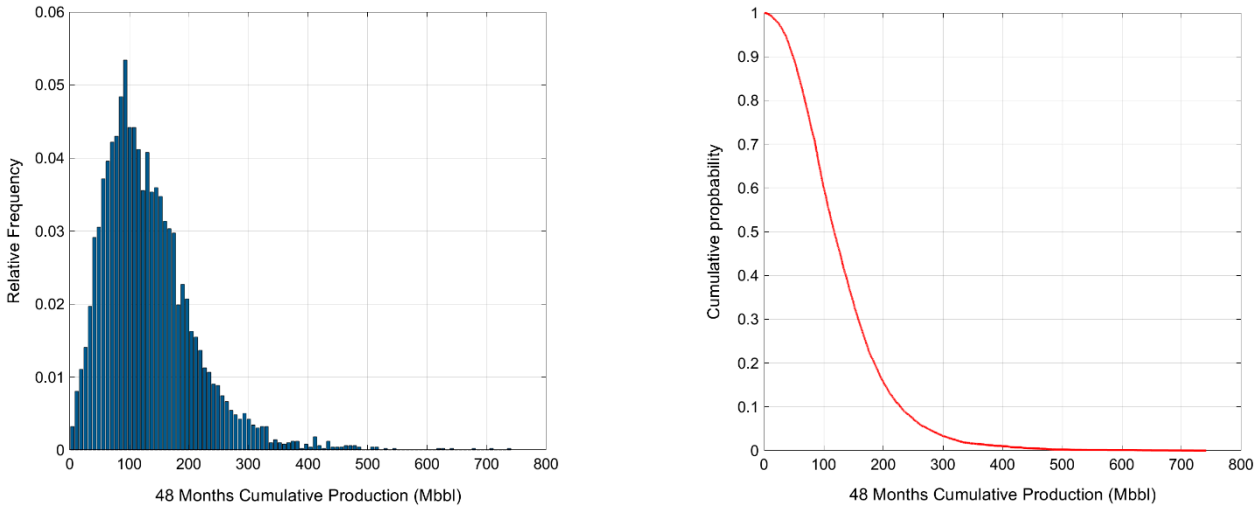


Figure 25. Prediction of the probabilistic production for new wells according to the production period: (**a**) NP 6; (**b**) NP 12; (**c**) NP 24.

These results indirectly confirmed the initial decline in the production at 6 and 12 months. The production behavior before the latter half period, which continues with low production, can be identified from the cumulative production of 48 months. By deriving these probabilistic results, uncertainty about the deterministic results was considered, and the cumulative production per unit length can be predicted using the cumulative production at a specific time, hydraulic fracturing design factors, and decline curve factors for each group.

To compare the reliability of the prediction results for new wells, we estimated the production using the commonly utilized hyperbolic function of Arps and Monte Carlo simulations. In this case, the analysis procedure proposed by Shin et al. [41] was utilized, and the results were derived by applying a kernel density function that could reflect the characteristics of the data in selecting the probability distribution shape for the decline curve factor of the entire well. Based on this probability distribution, we applied Monte Carlo simulations to estimate the cumulative production at 48 months for existing production wells, as shown in Figure 26.



(**a**) (**b**)

Figure 26. Probabilistic results of the total wells using the Arps hyperbolic for 48 months: (**a**) Histogram; (**b**) the cumulative probability distribution.

Therefore, for the entire well, P10, P50, and P90 were estimated to be 228.80, 117.34, and 48.69 Mbbl, respectively. To obtain the production per unit length, each of the 44 blind wells, assumed to be new wells, was divided into lateral lengths. For the cumulative production per unit length of each well, we identified that AAPE showed a relatively higher error than probabilistic results, with an error percentage of approximately 17% through values with the least error among the *p*-values as mentioned in the study. Accordingly, the probabilistic results for the existing multi-wells in the study area can be identified as large ranges of production prediction, such as the type curve, of new wells. In addition, the probabilistic production range may suggest the results of well units in a narrower range than the type curve.

6. Discussion

This study identified key factors in production-related attributes targeting 220 multi-wells produced over 60 months in the Eagle Ford shale. PCA was performed to extract features from various production-related attributes, and it was found that PC 1 was correlated with norm spacing, NVF, and NVP; TVD and HF stages for PC 2; and HF stage and NVP for PC 3. By classifying the groups, the production characteristics of the effects of hydraulic fracturing design factors were analyzed, and Cluster 2 conditions were identified as suitable when comparing the key factors with the productivity.

The production relationships of the classified groups were used to forecast the cumulative production of the existing wells. For each classified group, a trend line of NP 6 and NP 12 was drawn for NP 48, where the former case exhibited an error rate of approximately 10% for each group, and the latter case had an error rate of less than 10%. The trend line was derived by the analysis of production characteristics using the identified key factors of the existing wells; we could estimate the cumulative production per unit length using this

approach. In other words, the classification of production characteristics can be changed according to the changes in input attributes. As this study was found to be applicable even with very limited attributes, it can be applied to any field of shale formations. This could be achieved by using the three attributes of the new well (well locations, TVD) and four attributes (average value of spacing, probability value of dynamic parameters) obtained from the existing wells with a radius of 2500 ft or less of the new well.

Furthermore, we performed the production forecasting of new multi-wells using the existing wells. In the new wells with limited available field data, groups with different production characteristics were classified using only the key factors; machine learning with high accuracy or appropriate algorithms was identified based on the attributes and data features included. The accuracy of group classification can be improved by considering various input attributes in the future.

Owing to the characteristics of the shale reservoir with high production variability, the production trend can rapidly decrease after the initial production. To consider this, a prediction model of cumulative production per unit length of 6 and 12 months and 4 years was developed. The ANN model developed with the well location, TVD, hydraulic fracturing design factors, and dynamic parameters consisted of 80% training and 20% testing data. The results of verification with existing wells showed accuracy and error rates of approximately 98% and 2–3%, respectively. For the validated model, the cumulative production per unit length was predicted using probabilistic values of spacing and dynamic parameters, resulting in an accuracy exceeding 80% and an error of approximately 10%. In addition, the actual production value was included in the range of the predicted probabilistic values, and the error was small compared to the previous method.

The proposed multi-wells productivity analysis is a trial-and-error process based on data-driven analytics. It can be used to predict shale production to evaluate the economic feasibility of a project and establish a field development plan. Furthermore, this analysis can help to decrease the time and resources expended for data acquisition, thereby improving the reliability of productivity forecasts in shale development.

Author Contributions: Conceptualization, H.-J.S. and J.-S.L.; methodology, H.-J.S. and J.-S.L.; software, H.-J.S.; validation, H.-J.S. and J.-S.L.; formal analysis, H.-J.S., J.-S.L. and I.-S.J.; investigation, H.-J.S. and J.-S.L.; resources, J.-S.L.; data curation, H.-J.S. and J.-S.L.; writing—original draft preparation, H.-J.S.; writing—review and editing, J.-S.L. and I.-S.J.; visualization, H.-J.S.; supervision, J.-S.L.; project administration, H.-J.S.; funding acquisition, J.-S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korean government (MOTIE) (No. 20216110100050, Small and Medium Gas Field Exploitation, Production, and Field Operation Technique Development linked with small-scale gas power plants).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. U.S. Energy Information Administration (EIA). Annual Energy Outlook 2021 with Projections to 2050. Available online: https://www.eia.gov/outlooks/aeo/pdf/00%20AEO2021%20Chart%20Library.pdf (accessed on 12 March 2021).
2. Cao, Q.; Banerjee, S.; Gupta, S.; Li, J.; Zhou, W.; Jeyachandra, B. Data Driven Production Forecasting using Machine Learning. In Proceedings of the SPE Argentina Exploration and Production of Unconventional Resources Symposium, Buenos Aires, Argentina, 1–3 June 2016; SPE-180984-MS.
3. Suhag, A.; Ranjith, R.; Aminzadeh, F. Comparison of Shale Oil Production Forecasting using Empirical Methods and Artificial Neural Networks. In Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, TX, USA, 9–11 October 2017; SPE-187112-MS.

4. Sun, X.; Ma, X.; Kazi, M. Comparison of Decline Curve Analysis DCA with Recursive Neural Networks RNN for Production Forecast of Multi Wells. In Proceedings of the SPE Western Regional Meeting, Garden Grove, CA, USA, 22–27 April 2018; SPE-190104-MS.
5. Klie, H.; Florez, H. Data-Driven Discovery of Unconventional Shale Reservoir Dynamics. In Proceedings of the SPE Reservoir Simulation Conference, Galveston, TX, USA, 10–11 April 2019; SPE-193904-MS.
6. Mohaghegh, S.D. *Shale Analytics: Data-Driven Analytics in Unconventional Resources*, 1st ed.; Springer: Berlin, Germany, 2017; ISBN 978-3319487519.
7. Gaurav, A. Horizontal Shale Well EUR Determination Integrating Geology, Machine Learning, Pattern recognition and Multi-Variate Statistics Focused on the Permian Basin. In Proceedings of the SPE Liquids-Rich Basins Conference—North America, Midland, TX, USA, 13–14 September 2017; SPE-187494-MS.
8. Alabboodi, M.J.; Mohaghegh, S.D. Conditioning the Estimating Ultimate Recovery of Shale Wells to Reservoir and Completion Parameters. In Proceedings of the SPE Eastern Regional Meeting, Canton, OH, USA, 13–15 September 2016; SPE-184064-MS.
9. Li, Y.; Han, Y. Decline Curve Analysis for Production Forecasting Based on Machine Learning. In Proceedings of the SPE Symposium: Production Enhancement and Cost Optimisation, Kuala Lumpur, Malaysia, 7–8 November 2017; SPE-189205-MS.
10. He, Q. Smart Determination of Estimated Ultimate Recovery in Shale Gas Reservoir. In Proceedings of the SPE Eastern Regional Meeting, Lexington, KY, USA, 4–6 October 2017; SPE-187514-MS.
11. Vyas, A.; Gupta, A.D.; Mishra, S. Modeling Early Time Rate Decline in Unconventional Reservoirs Using Machine Learning Techniques. In Proceedings of the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, United Arab Emirates, 13–16 November 2017; SPE-188231-MS.
12. Bowie, B. Machine Learning Applied to Optimize Duvernay Well Performance. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AB, Canada, 13–14 March 2018; SPE-189823-MS.
13. BuKhamseen, N.Y.; Ertekin, T. Validating Hydraulic Fracturing Properties in Reservoir Simulation Using Artificial Neural Networks. In Proceedings of the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 24–27 April 2017; SPE-188093-MS.
14. Tandon, S. Integrating Machine Learning in Identifying Sweet Spots in Unconventional Formations. In Proceedings of the SPE Western Regional Meeting, San Jose, CA, USA, 23–26 April 2019; SPE-195344-MS.
15. Mohaghegh, S.D.; Gaskari, R.; Maysami, M. Shale Analytics: Making Production and Operational Decisions Based on Facts: A Case Study in Marcellus Shale. In Proceedings of the SPE Hydraulic Fracturing Technology Conference and Exhibition, The Woodlands, TX, USA, 24–26 January 2017; SPE-184822-MS.
16. Amr, S.; El Ashhab, H.; El-Saban, M.; Schietinger, P.; Caile, C.; Kaheel, A.; Rodriguez, L. A Large-Scale Study for a Multi-Basin Machine Learning Model Predicting Horizontal Well Production. In Proceedings of the SPE Annual Technical Conference and Exhibition, Dallas, TX, USA, 24–26 September 2018; SPE-191538-MS.
17. Grechka, V.; Li, Z.; Howell, B.; Garcia, H.; Wooltorton, T. Microseismic imaging of unconventional reservoirs. In Proceedings of the 2018 SEG International Exposition and Annual Meeting, Anaheim, CA, USA, 14–19 October 2018; SEG-2018-2995627.
18. Hetz, G.; Kim, H.; Datta-Gupta, A.; King, M.J.; Przybysz-Jarnut, J.K.; Lopez, J.L.; Vasco, D. History Matching of Frequent Seismic Surveys Using Seismic Onset Times at the Peace River Field, Canada. In Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, TX, USA, 9–11 October 2017; SPE-187310-MS.
19. Baek, S.; Akkutlu, I.Y.; Lu, B.; Ding, S.; Xia, W. Shale Gas Well Production Optimization using Modified RTA Method—Prediction of the Life of a Well. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, Denver, CO, USA, 22–24 July 2019; URTEC-2019-185.
20. Economides, M.J.; Nolte, K.G. *Reservoir Stimulation*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2000; ISBN 978-0471491927.
21. Shin, H.D.; Lee, K.S.; Jang, I.S.; Park, C. *Unconventional Resources Development*; CIR: Seoul, Korea, 2015; ISBN 979-1156101765.
22. Wigwe, M.E.; Bougre, E.S.; Watson, M.C.; Giussani, A. Comparative evaluation of multi-basin production performance and application of spatio-temporal models for unconventional oil and gas production prediction. *J. Petrol. Explor. Prod. Technol.* **2020**, *10*, 3091–3110. [CrossRef]
23. Ma, Y.Z.; Holditch, S. *Unconventional Oil and Gas Resources Handbook: Evaluation and Development*, 1st ed.; Gulf Professional Publishing: Oxford, UK, 2015; ISBN 978-0128022382.
24. Wigwe, M.E.; Bougre, E.S.; Watson, M.C.; Giussani, A. Spatio-temporal Models for Big Data and Applications on Unconven-tional Production Evaluation. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, Austin, TX, USA, 20–22 July 2020; URTEC-2020-2855.
25. Satter, A.; Iqbal, G.M. *Reservoir Engineering: The Fundamentals, Simulation, and Management of Conventional and Unconventional Recoveries*, 1st ed.; Gulf Professional Publishing: Oxford, UK, 2015; ISBN 978-0128002193.
26. James, S. *Shale Gas Production Processes*, 1st ed.; Gulf Professional Publishing: Oxford, UK, 2013; ISBN 978-0124045712.
27. Texas RRC. Eagle Ford Shale Information. Available online: www.rrc.texas.gov/oil-and-gas/major-oil-and-gas-formations/eagle-ford-shale/ (accessed on 10 August 2021).
28. Arps, J.J. Analysis of Decline Curves. *Trans. AIME* **1945**, *160*, 228–247. [CrossRef]
29. Seshadri, J.N.; Mattar, L. Comparison of Power Law and Modified Hyperbolic Decline Methods. In Proceedings of the Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, Canada, 19–21 October 2010; SPE-137320-MS.

30. Clark, A.J.; Lake, L.W.; Patzek, T.W. Production Forecasting with Logistic Growth Models. In Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, CO, USA, 30 October–2 November 2011; SPE-144790-MS.
31. Yu, S.; Miocevic, D.J. An Improved Method to Obtain Reliable Production and EUR Prediction for Wells with Short Production History in Tight/Shale Reservoirs. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, Denver, CO, USA, 12–14 August 2013; URTEC-1563140-MS.
32. Joshi, K.; Lee, W.J. Comparison of Various Deterministic Forecasting Techniques in Shale Gas Reservoirs. In Proceedings of the SPE Hydraulic Fracturing Technology Conference, The Woodlands, TX, USA, 4–6 February 2013; SPE-163870-MS.
33. Shin, H.J.; Kim, J.S.; Lim, J.S. Application of various deterministic decline curve analyses in resource plays. *Geosyst. Eng.* **2015**, *18*, 61–72. [CrossRef]
34. Kang, P.S.; Shin, H.J.; Lim, J.S. Effect of Shale Reservoir Property and Condition of Hydraulic Fracture on Decline Curve Analysis Factor. *J. Korean Soc. Miner. Energy Resour. Eng.* **2017**, *54*, 223–232. [CrossRef]
35. Kurtoglu, B.; Cox, S.A.; Kazemi, H. Evaluation of Long-Term Performance of Oil Wells in Elm Coulee Field. In Proceedings of the Canadian Unconventional Resources Conference, Calgary, AB, Canada, 15–17 November 2011; SPE-149273-MS.
36. Meyet Me Ndong, M.P.; Dutta, R.; Burns, C. Comparison of Decline Curve Analysis Methods with Analytical Models in Unconventional Plays. In Proceedings of the SPE Annual Technical Conference and Exhibition, New Orleans, LA, USA, 30 September–2 October 2013; SPE-166365-MS.
37. Rezaee, R. *Fundamentals of Gas Shale Reservoirs*, 1st ed.; Wiley: Hoboken, NJ, USA, 2015; ISBN 978-1118645796.
38. Shin, H.J.; Lim, J.S.; Shin, S.H. Estimated ultimate recovery prediction using oil and gas production decline curve analysis and cash flow analysis for resource play. *Geosyst. Eng.* **2014**, *17*, 78–87. [CrossRef]
39. Gonzalez, R.; Gong, X.; McVay, M. Probabilistic Decline Curve Analysis Reliably Quantifies Uncertainty in Shale Gas Reserves Regardless of Stage of Depletion. In Proceedings of the SPE Eastern Regional Meeting, Lexington, KY, USA, 3–5 October 2012; SPE-161300-MS.
40. Kim, J.S.; Shin, H.J.; Lim, J.S. Probabilistic Decline Curve Analysis for Forecasting Estimated Ultimate Recovery in Shale Gas Play. *J. Korean Soc. Miner. Energy Resour. Eng.* **2014**, *51*, 808–819. [CrossRef]
41. Shin, H.J.; Hwang, J.Y.; Lim, J.S. Probabilistic Prediction of Estimated Ultimate Recovery in Shale Reservoir using Kernel Density Function. *J. Korean Inst. Gas* **2017**, *21*, 61–69. [CrossRef]
42. Luo, G.; Tian, Y.; Bychina, M.; Ehlig-Economides, C. Production Optimization Using Machine Learning in Bakken Shale. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, TX, USA, 23–25 July 2018; URTEC-2902505-MS.
43. U.S. Energy Information Administration (EIA). The Number of Drilled but Uncompleted Wells in the United States Continues to Climb. Available online: https://www.eia.gov/todayinenergy/detail.php?id=39332 (accessed on 29 April 2021).
44. Perrier, S.; Delpeint, A.; Shawutii, Z.; Shrestha, A. Machine-Learning Based Analytics Applied to Stimulation Performance in the Utica Shale: Case Study and Lessons Learned. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, TX, USA, 23–25 July 2018; URTEC-2902017-MS.
45. Chen, C.; Gao, G.; Gelderblom, P.; Jimenez, E. Integration of cumulative-distribution-function mapping with principal-component analysis for the history matching of channelized reservoirs. *SPE Reserv. Eval. Eng.* **2016**, *19*, 278–293. [CrossRef]
46. Lim, J.S.; Kang, J.M.; Kim, J. Multivariate Statistical Analysis for Automatic Electrofacies Determination from Well Log Measurements. In Proceedings of the SPE Asia Pacific Oil and Gas Conference, Kuala Lumpur, Malaysia, 14–16 April 1997; SPE-38028-MS.
47. Yoo, H.J.; Lim, J.S.; Kim, S.J. Electrofacies determination from well logs using fuzzy clustering analysis. *J. Korean Soc. Miner. Energy Resour. Eng.* **2009**, *46*, 424–430.
48. Tan, P.; Steinbach, M.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*, 2nd ed.; Pearson: London, UK, 2018; ISBN 978-0133128901.
49. Zaefferer, M. Optimization and Empirical Analysis of an Event Detection Software for Water Quality Monitoring. Master's Thesis, Technical University of Cologne, Cologne, Germany, June 2012.
50. MathWorks. Machine Learning. Available online: https://kr.mathworks.com/campaigns/offers/machine-learning-with-matlab.html (accessed on 10 August 2021).
51. Politecnico di Milano. A Tutorial on Clustering Algorithms. Available online: https://matteucci.faculty.polimi.it/Clustering/tutorial_html/index.html (accessed on 11 August 2021).
52. Harwood, C.; Wipat, A. *Methods in Microbiology*; Academic Press: Cambridge, MA, USA, 2012; Volume 39, ISBN 978-0080993874.
53. Theodoridi, S.; Koutroumbas, K. *Pattern Recognition*, 4th ed.; Academic Press: Cambridge, MA, USA, 2008; ISBN 978-0123744913.
54. Angelov, P.P.; Gu, X. *Empirical Approach to Machine Learning*; Springer: Berlin, Germany, 2019; ISBN 978-3030023836.
55. Cunningham, P.; Delany, S.J. *K-Nearest Neighbour Classifiers*; Technical Report UCD-CSI-2007-4; School of Computer Science & Informatics, University College Dublin: Dublin, Ireland, 2007.
56. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin, Germany, 2009; ISBN 978-0387848570.
57. Ertekin, T.; Silpngarmlers, N. Optimization of formation analysis and evaluation protocols using neuro-simulation. *J. Petrol. Sci. Eng.* **2005**, *49*, 97–109. [CrossRef]