Article

Identification of the Thief Zone Using a Support Vector Machine Method

Cheng Fu [1,2,*], Tianyue Guo [1], Chongjiang Liu [3], Ying Wang [4] and Bin Huang [1,*]

1   Key Laboratory of Enhanced Oil Recovery, Ministry of Education, College of Petroleum Engineering, Northeast Petroleum University, Daqing 163318, China; yueyuepikaqiu@163.com
2   Post-Doctoral Scientific Research Station, Daqing Oilfield Company, Daqing 163413, China
3   Research Institute of Production Engineering, Daqing Oilfield, Daqing 163453, China; liuchongjiang@petrochina.com.cn
4   Aramco Asia, Beijing 100102, China; ying.wang.1@aramcoasia.com
*   Correspondence: cheng_fu111@163.com (C.F.); huangbin111@163.com (B.H.)

check for updates

Abstract: Waterflooding is less effective at expanding reservoir production due to interwell thief zones. The thief zones may form during high water cut periods in the case of interconnected injectors and producers or lead to a total loss of injector fluid. We propose to identify the thief zone by using a support vector machine method. Considering the geological factors and development factors of the formation of the thief zone, the signal-to-noise ratio and correlation analysis method were used to select the relevant evaluation indices of the thief zone. The selected evaluation indices of the thief zone were taken as the input of the support vector machine model, and the corresponding recognition results of the thief zone were taken as the output of the support vector machine model. Through the training and learning of sample sets, the response relationship between thief zone and evaluation indices was determined. This method was used to identify 82 well groups in M oilfield, and the identification results were verified by a tracer monitoring method. The total identification accuracy was 89.02%, the positive sample identification accuracy was 92%, and the negative sample identification accuracy was 84.375%. The identification method easily obtains data, is easy to operate, has high identification accuracy, and can provide certain reference value for the formulation of profile control and water shutoff schemes in high water cut periods of oil reservoirs.

Keywords: thief zone; support vector machine; signal-to-noise ratio; correlation analysis; tracer monitoring

1. Introduction

The earliest research on "thief zone" can be tracked to the 1950s [1]. A thief zone refers to the low-resistivity seepage channel formed locally in the reservoir due to geological factors (including porosity, permeability, and effective thickness, etc.) and development factors (including cumulative water injection, daily water injection, water injection pressure, bottom hole pressure, water cut, cumulative liquid production, and daily liquid production, etc.) [2]. During the oil field development, the formation pressure gradually decreases with the exploitation of the oilfield. In order to maintain the formation pressure and extend the life of the oilfield, secondary oil recovery is generally adopted to develop the oilfield by water injection. At a later stage of water flooding development, the injected water forms an obvious thief zone along this channel, resulting in a large amount of injected water creating an invalid cycle, which greatly reduces the development effect and economic benefits of the oilfield [3]. Therefore, it is very important to identify the well group of the thief zone for profile control and water plugging.

At present, there are many methods used to identify thief zones, each with its own principle and basis. The commonly used methods include well logging, coring, well testing, dynamic monitoring between wells [4,5], and reservoir engineering methods combined with modern calculation methods, including math or computer.

In 2002, Wang et al. identified the large thief zone by using water injection profile logging data [6]. In 2007, Meng et al. proposed the application of conventional logging data and Fisher discriminant to identify the thief zone [7,8]. In 2007, Al-Dhafeeri et al. identified the thief zone using core data and a production logging test [9]. In 2008, Li et al. proposed that the thief zone could be identified by comparing the isotopic attenuation amplitude traced in different time periods by using a well log isotope curve [10]. In 2008, Li et al. described a method combining production logging test (PLT), nuclear magnetic resonance (NMR), and high-resolution image logging to identify the thief zone [11]. In 2008, Chen et al. studied thief zones by using PLTs and summarized the distribution of different types of thief zones [12]. In 2013, John et al. used distributed temperature sensing (DTS) technology combined with PLTs and water flow logging (WFL) to detect the location of the thief zone for the first time [13]. These logging methods are simple and easy to use to identify the thief zone. However, logging can only be performed in the immediate vicinity of the situation and requires both early and late logging data to be complete and can interfere with well performance during testing.

In 1998, He et al. proposed to identify the thief zone by analyzing the changes in core permeability [14]. In 2001, Liao et al. established a prediction model of pore throat volume distribution based on porosity and permeability by using a mercury injection curve and described the pore structure of the thief zones [15]. In 2002, based on Purcell capillary bundle model He et al. established a prediction model of pore throat volume distribution with the principle and method of geometry and described the pore structure in the reservoir [16]. All the above methods can accurately describe the thief zone, but a large amount of core data before and after the formation of the thief zone is needed, and the coring cost is too high.

In 2003, Shi et al. used typical fitting curves of well test data to identify thief zones, drew curves of measured points corresponding to different time periods, and established a theoretical model of well test for thief zones [17]. In 2005, Yin et al. identified the thief zone by analyzing the change of reservoir property in the process of water injection and the reservoir condition between the injection well and the production well [18]. In 2013, Feng et al. used pressure transient analysis to characterize the thief zone of mature water drive reservoirs and established a mathematical model of intersecting wells with high permeability bands [19]. In 2015, Liu proposed the use of dimensionless PI (Pressure Index) value to identify the thief zone and customized the bottom hole pressure and pressure derivative log curve well test interpretation chart [20]. This method eliminates the influence of reservoir pressure relief ability on the pressure drop curve, but it must be based on the ideal model, which is different from the actual formation. This is also the defect of all well testing methods for identifying thief zones.

In recent years, the tracer method proposed by Izgec, Kabir [21], Batycky [22], Wang et al. [23] to identify the thief zone has been proven to be accurate, but it has a long test time, high requirements for continuous detection, and a high cost.

At present, the mathematical evaluation method is widely used to identify the thief zone of waterflood sandstone reservoirs. Gray theory, a method that studies the phenomenon of information which is partly clear and partly unclear with uncertainty, was used to calculate the correlation between various factors to diagnose the existence of the thief zone by Dou in 2001 [24]. In 2002, Zeng et al., guided by the principle of percolation mechanics and reservoir engineering method, proposed a mathematical model to describe the thief zone, and identified and described the thief zone in the oil layer by using gray correlation theory and conventional dynamic data [25]. In 2003, Liu et al. applied the fuzzy discriminant method of expert system to track and identify thief zone [26].

Although these methods can identify the thief zone, they have the same problem: In the identification process, subjectivity is strong, and the judgment standard of the thief zone is difficult to determine, which is not conducive to practical application. In 2011, Wang et al. first used the

ISODATA clustering analysis method combined with different dynamic data characteristics of oil and water wells to determine whether there is a thief zone in each well and the level of the thief zone, which solved the rationality problem of threshold value well [27]. However, this method only describes the situation near the wellbore, and the selection index of evaluation is too artificial and lacks the corresponding selection method. In 2015, Ding et al. proposed a method to determine the thief zone by using automatic history matching [28] and fuzzy mathematics [29]. The method considers the uncertainty of geology, but the parameters are difficult to obtain. In 2018, Huang et al. used a multi-layer weighted principal component analysis method and multi-level fuzzy comprehensive evaluation method to identify the thief zone [30,31]. This method fully considers the influence of each evaluation index on the recognition result, effectively improves the recognition accuracy, but the recognition process is complex.

In order to improve the accuracy of thief zone identification and simplify the identification process, we use a method of support vector machine [32] for identification. As early as 2009, Jin used a support vector machine method combined with a logging curve to identify the thief zone [33], but his identification process was too simple, and the accuracy was not high. In 2015, Chen et al. used an improved support vector machine for quantitative identification of the thief zone [34]. He took the change of reservoir permeability as the identification standard. The identification results were not representative, and the index selection was too simple.

Based on the above research, we used the signal-to-noise ratio and correlation analysis to screen the indices affecting the thief zone to eliminate the influences of human factors on the screening indices. Then, the optimal kernel function was selected by a cross-validation method to make the model more reliable and stable, and the discriminant result more representative. Finally, the method was applied to identify thief zones of M oilfield, and the accuracy of identification was verified by tracer monitoring method.

2. Support Vector Machine Method to Identify the Thief Zone

In this paper, the identification of the thief zone is mainly divided into the following four steps: First, the signal-to-noise ratio (SNR) and correlation analysis method are used to extract and select the thief zone evaluation indices. Second, the evaluation indices are standardized. Third, the cross-validation method is used to select the optimal kernel function, and the optimal kernel function is used to establish a support vector machine model. And finally, the identification results of the support vector machine (SVM) model are verified and analyzed by tracer monitoring method. The entire thief zone identification process is represented by a flow chart, as shown in Figure 1.
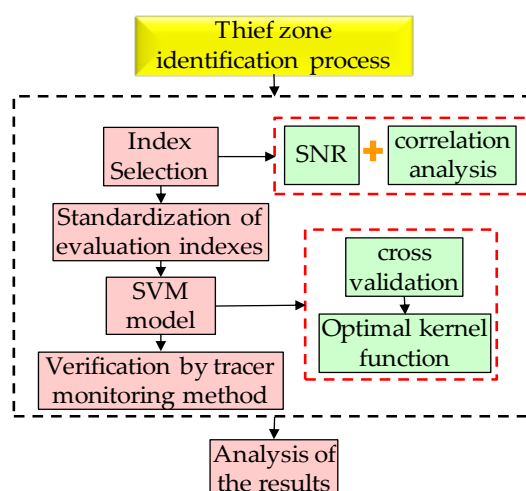

Figure 1. Flow chart of thief zone identification.

2.1. Extraction and Selection of Thief Zone Evaluation Indices

In the process of extracting evaluation indices (such as permeability, porosity and water cut etc.), on the one hand, it is necessary to ensure that each index has a certain influence on the formation of the thief zone. On the other hand, it is also necessary to consider the actual data of oilfields to select indices that are easy to obtain data. The selection of an evaluation index refers to the process of selecting the most effective index combination for classification and recognition through screening the extracted indices or some transformation. The extraction and selection of evaluation indices are very important in the identification of thief zones, which directly determines the accuracy of the identification results.

In order to avoid the problem that the classification accuracy of SVM is too low due to too many indices and the correlation among indices, we comprehensively use signal-to-noise ratio and correlation analysis method to screen indices.

2.1.1. Signal-to-Noise Ratio

The signal-to-noise ratio (SNR) is used to rank the different attributes according to each index. The noise index can be effectively eliminated so that the index in front contains more classification information. Using these indices to classify can achieve higher accuracy. The amount of classification information contained in each index is determined by the following formula.

$$SNR_i = \frac{\left| \mu_P(i) - \mu_N(i) \right|}{\sigma_P(i) + \sigma_N(i)} \tag{1}$$

where $\mu_P(i)$ and $\mu_N(i)$ represent the mean value of the ith feature of positive and negative samples respectively, $\sigma_P(i)$ and $\sigma_N(i)$ represent the standard deviation of the ith feature of positive and negative samples respectively, and SNR_i represents the signal-to-noise ratio of the ith feature. The larger the SNR_i, the more classified information this feature contains, and the more favorable it is for classification. The advantages of SNR are its simple calculation process and less time-consuming nature. The disadvantage is that the interaction between the indicators is not taken into account.

2.1.2. Correlation Analysis

Correlation analysis is an important means of finding the relationship between indices. The correlation coefficient is a statistical index of the degree of a close relationship between response variables. The range of correlation coefficient is between 1 and −1. 1, which means that two variables are completely linearly correlated, −1 means that two variables are completely negatively correlated, and 0 means that two variables are not correlated. The closer the correlation coefficient is to 0, the weaker the correlation is. The following is the formula for calculating the correlation coefficient.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \tag{2}$$

$$S_{xy} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n - 1} \tag{3}$$

$$S_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n - 1}} \tag{4}$$

$$S_y = \sqrt{\frac{\sum (y_i - \overline{y})^2}{n - 1}} \tag{5}$$

where r_{xy} is the sample correlation coefficient, S_{xy} is the sample covariance, S_x is the sample standard deviation of X, and S_y is the sample standard deviation of Y. The advantage of correlation analysis is that it can consider the interaction between indices.

2.2. Evaluation Index Standardization

In the multi-index evaluation system, due to the different nature of each evaluation index, there are usually different dimensions and quantities. When the level of each index is very different, if the original index value is used to analyze directly, the function of the higher value index in the comprehensive analysis will be highlighted, and the function of the lower value index will be relatively weakened. Therefore, in order to ensure the reliability of the results, it is necessary to standardize the original index data. At present, there are many methods of data standardization, which can be summed up as straight-line methods (such as the extreme value method or standard deviation method), broken line methods (such as the three-fold line method), and curve methods (such as semi-normal distribution). For simplicity and ease of calculation, this paper uses the "min–max standardization" method to standardize data.

$$x' = \frac{x - minA}{maxA - minA} \tag{6}$$

where x' is normalized value, x is sample value, max A is the maximum value of attribute A, min A is the minimum value of attribute A.

2.3. Support Vector Machine Model

As a classification method proposed by Vapnik et al., support vector machine [35] is mainly applied in the field of pattern recognition and has many unique advantages in small-sample, nonlinear, and high-dimensional pattern recognition. In recent years, it has been successfully applied in image recognition [36], signal processing [37], gene map recognition [38], and benign and malignant tumor recognition [39], showing its advantages. Moreover, field data are characterized by extremely complex relationships and few records, which is suitable for classification prediction with support vector machines.

The basic principle of support vector machine is to map the sample space to a feature space of high or even infinite dimensions (namely Hilbert space) through nonlinear mapping, and to find the optimal classification surface in this feature space. Then, the nonlinear separable problem in the sample space can be transformed into a linear separable problem in the feature space by taking appropriate kernel functions.

For example, set the training sample set as $T = \{(x_1, y_1), \ldots, (x_i, y_i)\}, x_i \in R^n, y_i \in (+1, -1),$ $i = 1, \ldots, l$, where l is the number of observed samples. If there is a hyperplane equation $w * x + b = 0$ to make the sample set satisfy $y_i(w * x_i + b) \geq 1$, the training set is called linearly separable. Figure 2 shows the basic idea of SVM, in which the circular point and the square point represent two kinds of samples respectively. N is the classification hyperplane. N1 and N2 are the planes parallel to the classification hyperplane respectively. The samples on them are closest to the classification hyperplane, and the distance between them is the classification interval. The so-called optimal classification hyperplane requires that the classification plane not only correctly separates the two categories, but also maximizes the classification interval, which is $2/\|w\|$. The nearest vector to the optimal hyperplane is the support vector, and the distance between the support vector and the hyperplane is $1/\|w\|$. Thus, the problem of constructing the optimal hyperplane is transformed into the problem of seeking the optimal solution of quadratic programming.

The objective function is:

$$\min_{w,b} \left[\frac{1}{2} \|w\|^2 \right]. \tag{7}$$

The restrictions are:

$$\text{s.t.} \begin{cases} y_i((w \cdot x_i) + b) \geq 1, \ i = 1, 2, \ldots, l \\ y_i \in \{-1, 1\} \end{cases} \tag{8}$$
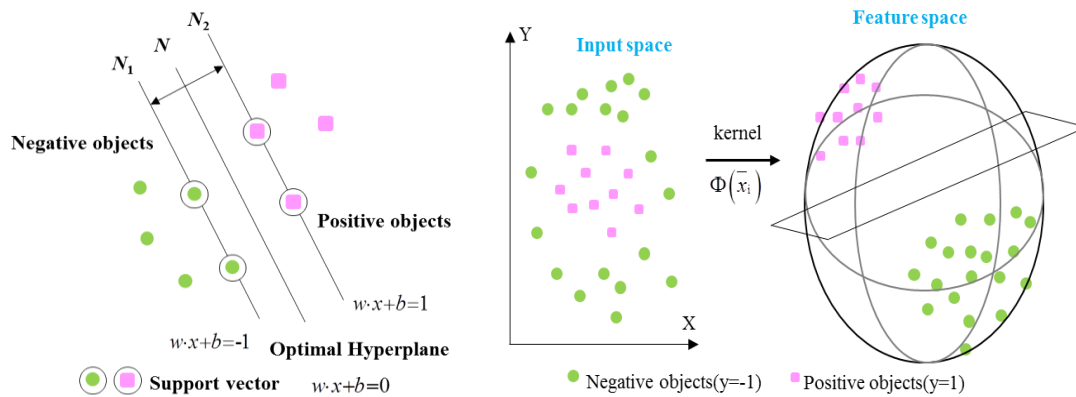
Figure 2. Schematic diagram of support vector machine method.

The optimal classification function obtained after solving the above problems is

$$f(x) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i^* y_i(x_i \cdot x) + b*\right). \tag{9}$$

When the training sample is linearly indivisible, slack variables $\xi = (\xi_1, \ldots, \xi_l)^T$ need to be introduced. Then, SVM can be expressed as the following quadratic optimization problem, where $C > 0$ is the penalty parameter.

$$\min_{w,b}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \zeta_i\right] \tag{10}$$

$$\text{s.t.}\begin{cases} y_i((w \cdot x_i) + b) \geq 1 - \zeta_i, \ i = 1, 2, \ldots, l, \\ \zeta_i \geq 0, \ i = 1, 2, \ldots l, \end{cases} \tag{11}$$

To solve this quadratic optimization problem, the kernel function $k(x_i, y_i)$ needs to be introduced, and the expression of SVM can be obtained as follows:

$$f(x) = \text{sign}\left[\sum_{i=1}^{l} \alpha_i y_i k(x_i \cdot x) + b\right]. \tag{12}$$

For the nonlinear problem, the input space needs to be mapped to a high-dimensional feature space by the kernel function $k(x_i, y_i)$, as shown in Figure 2. Because the dual problem above only involves the inner product operation between training samples, as long as the inner product of the kernel function $k(x_i, y_i)$ is equal to the inner product of X and Y mapped in the high-dimensional feature space, a support vector machine can be determined by the kernel function $k(x_i, y_i)$, so as to avoid the dimension disaster problem in the feature space.

In general, the four kernel functions include the following:

(1)   The linear kernel function $K(x, y) = x \cdot y$;
(2)   The binomial kernel function $K(x, y) = [(x \cdot y) + 1]d$;
(3)   The radial basis kernel function (RBF) $K(x, y) = \exp(-|x - y|^2/d^2)$; and
(4)   The sigmoid kernel function (neural network) $K(x, y) = \tanh(a(x \cdot y) + b)$.

Each of the four functions has its own characteristics, as shown in Table 1 below.

Table 1. Comparison of kernel functions.

Kernel Function	Characteristics	Applicable Situation	Advantage
Linear kernel function	Simple kernel	Mainly for the linearly separable case	Fewer parameters and faster
Polynomial kernel function	Global kernel	Suitable for orthogonal normalization data, more parameters are needed	The values of kernel functions are allowed to be affected by data points that are far apart
Radial basis kernel function (RBF)	Local kernel function	Small sample, large sample, high-dimensional, low-dimensional are applicable	It has good anti-interference ability, strong locality, and few parameters
Sigmoid kernel function	Derived from neural networks	For some parameters, sigmoid and RBF have similar performance	Good generalization ability for unknown samples

In order to select the optimal kernel function, we adopted a five times cross validation method (that is, the sample database data set is randomly divided into five parts, each part is taken as a test set in turn, and the rest is taken as a training set) for training and testing. *Sen*, *Spe*, and *Q* were used for statistical analysis:

$$Sen = TP/(TP + FN) \tag{13}$$

$$Spe = TN/(TN + FP) \tag{14}$$

$$Q = (TP + TN)/(TP + FN + TN + FP). \tag{15}$$

where the positive samples are non-thief zone well groups known in the oilfield. Negative samples are thief zone well groups known in the oilfield. True positive *TP* indicates the number of samples accurately judged as positive samples in the test set. False negative *FN* indicates the number of samples wrongly judged as negative in the test set. True negative *TN* indicates the number of samples accurately judged as negative samples in the test set, and false positive *FP* indicates the number of samples wrongly judged as positive in the test set. *Sen* is the sensitivity of positive samples in the test set, and *Spe* is the specificity of negative samples in the test set. The larger the *Sen*, the stronger the recognition ability of the positive sample. The larger the *Spe*, the better the discrimination effect on negative samples. *Q* is the total test accuracy.

Finally, the kernel function with the highest total test accuracy and the strongest sensitivity to negative samples (i.e., the thief zone) that satisfies the applicable situation should be selected as the optimal kernel function. Then, the selected thief zone evaluation indices are taken as the input, and the corresponding thief zone recognition results are taken as the output. By using the optimized kernel function and through continuous training and learning, the trained SVM model can be used as a thief zone recognition model.

3. Examples of Application

Taking a block of M oil field as an example, based on the existing data of the oil field, the thief zone identification was carried out according to the above method, and the identification results were verified by tracer monitoring method.

3.1. Oilfield Overview

M oilfield is a heterogeneous sandstone reservoir with positive rhythm deposition, with a development area of 9.56 km^2 and geological reserves of 1556.51×10^4 t. The top burial depth of the oil layer is from 751 to 1047 m, and the oil–water interface depth is from 1088 to 1197 m. The reservoir's original formation pressure was 11.07 Mpa. The average air permeability of this block in M oilfield is 632×10^{-3} μm^2, porosity is 27.4%, effective thickness is 43.5 m, reservoir temperature is 42.7~51 °C, and underground crude oil density is 0.89 g/cm^3.

After more than 20 years of water flooding development, the comprehensive water cut of M oilfield has reached 82%. Long-term water flooding development leads to serious development of thief zones in this oilfield, and a large amount of injected water rushes along the thief zones, resulting in inefficient or even invalid circulation of injected water in many blocks, which greatly reduces the oilfield development effect and economic benefits. Therefore, strengthening the identification of thief zones has become a prominent task.

3.2. Evaluation Indices

According to the extraction principles of the above evaluation indices and considering the geological factors and development factors that affect the formation of thief zones, the indices, such as permeability K, porosity ϕ, effective thickness h, cumulative water injection W_i, daily water injection q_i, water injection pressure P_{wo}, bottom hole pressure P_{wf}, water cut f_w, cumulative liquid production W_l, and daily liquid production q_l were selected.

By calculating the signal-to-noise ratio (SNR) and correlation coefficient, the contribution of each index to classification, and the correlation among the indices were obtained simultaneously, as shown in Tables 2 and 3.

Table 2. Indices ranking results of signal-to-noise ratio (SNR).

Number	1	2	3	4	5	6	7	8	9	10
Index	q_i	h	W_i	ϕ	W_q	K	f_w	P_{wf}	q_l	P_{wo}
SNR	0.32	0.24	0.23	0.22	0.22	0.19	0.17	0.12	0.06	0.03

Table 3. Relevance among indices.

Index	Relevance Ranking				
	1	2	3	4	5
q_i	h	ϕ	W_i	q_l	W_q
q_l	W_i	q_i	h	ϕ	
h	W_i	q_i	ϕ	q_l	W_q
ϕ	h	q_i	W_i	q_l	K
W_i	h	q_l	q_i	ϕ	
W_q	h	q_i			

Numbers 1–10 in Table 2 indicate the contribution degree of each index to classification: 1 indicates the largest contribution degree, and 10 indicates the smallest contribution degree, which decreases in turn. In Table 2, it can be seen that the SNR of water injection pressure was 0.03, which contributed little to the classification, followed by daily liquid production and bottom hole pressure. The maximum contribution of daily water injection to classification was 0.32. Considering the result of the SNR calculation, the main objects of screening were water injection pressure, daily liquid production, and bottom hole pressure.

Numbers 1–5 in Table 3 indicate the degree of correlation among the indices: 1 indicates the strongest degree of correlation, and 5 indicates the weakest degree of correlation, which decreases in turn. The indices listed in the table are those whose correlation coefficient is greater than 0.5. Those whose correlation coefficient is less than 0.5 have a weak correlation and were not considered for the time being. In Table 3, we can clearly see the correlation among the indices. Considering the results of correlation calculation, daily water injection, daily liquid production, effective thickness, porosity, and cumulative water injection were the main targets for screening.

According to Tables 2 and 3, from the perspective of oilfield development, the permeability, porosity, and effective thickness are the basic indices of the oilfield, and the greater the permeability, porosity, and effective thickness, the more obvious the heterogeneity of the reservoir, and the easier it is to form the thief zone. Therefore, these indices should be retained. Because daily water injection

and daily liquid production, cumulative water injection and cumulative liquid production, and water injection pressure and bottom hole pressure are three pairs of one-to-one corresponding indices, and daily water injection also had a great impact on the cumulative water injection and water injection pressure, in the screening process, only one pair of indices with the largest contribution could be retained. The SNR of daily water injection was the highest, and its contribution to classification was the highest. However, it also had a high correlation with other indices, which may lead to low classification accuracy. In addition, the SNR of daily liquid production was 0.06, which means it makes a low contribution to classification. Therefore, the pair of indices of daily water injection and daily liquid production was ignored. In addition, because the SNRs of water injection pressure and bottom hole pressure were relatively low, they did not contribute much to classification. Therefore, the pair of indices of water injection pressure and bottom hole pressure was also ignored. Thus, the indices of cumulative water injection and cumulative liquid production were retained.

In summary, the optimal evaluation index system is shown in Table 4.

Table 4. Evaluation index system.

No.	Index	Unit
1	Permeability	μm^2
2	Porosity	%
3	Effective thickness	m
4	Cumulative water injection	m^3
5	Water cut	%
6	Cumulative liquid production	m^3

3.3. SVM Identification Method

According to the actual working conditions of M oilfield, after years of verification by field experts and actual statistical data of the oil field, 20 groups of typical well groups with highly developed thief zones and 20 groups of ordinary well groups with good production conditions were obtained. The 40 groups of well groups were taken as the training samples for discrimination. The sample database of production data after statistical standardization is shown in Table 5.

Table 5. Sample database.

No.	h	ϕ	k	W_i	f_w	W_q	Identify Results
1	0.0000	0.0000	0.0476	0.0082	0.5955	0.0024	−1
2	0.0000	0.0000	0.0000	0.0195	0.9101	0.0335	−1
3	0.3077	0.8000	1.0000	0.1377	0.8315	0.0292	−1
4	0.0000	0.0000	0.0476	0.0145	0.7303	0.0020	−1
5	0.1538	0.6000	0.2690	0.1178	0.3596	0.0419	−1
6	0.3077	0.8000	0.0476	0.3663	0.6854	0.2453	−1
7	0.0000	0.0000	0.0476	0.0489	0.1348	0.0009	−1
8	0.2308	0.6000	0.4405	0.1667	0.6180	0.0363	−1
9	0.9231	0.8000	0.6071	0.6103	0.9775	1.0000	−1
10	0.0000	0.0000	0.0000	0.0221	0.7978	0.0013	−1
11	0.2308	0.4000	0.0714	0.1377	0.5169	0.0347	1
12	0.4615	0.8000	0.0833	0.5093	0.8315	0.1396	1
13	0.2308	0.4000	0.0714	0.2780	0.9101	0.0779	1
14	0.2308	0.4000	0.0714	0.2780	0.8315	0.0791	1
15	0.0000	0.0000	0.0000	0.0467	0.9101	0.0183	1
16	0.1538	0.4000	0.0476	0.1574	0.4382	0.0071	1
17	0.6154	0.8000	0.5000	0.5329	1.0000	0.0029	1
18	0.1538	0.4000	0.0476	0.0952	0.4831	0.0055	1
19	0.3846	0.8000	0.0595	0.2208	0.4831	0.0126	1
20	0.0000	0.0000	0.0476	0.0115	0.4831	0.0008	1

In order to select the kernel function with the best recognition effect, five cross-validations were carried out to calculate *Sen*, *Spe*, and *Q*. The results are shown in Figure 3.

It can be seen in Figures 3–6 that the classification accuracy of linear kernel function and radial basis kernel function was 90.5%, the polynomial kernel function was 80%, and the Sigmoid kernel function was 74% when the above four kernel functions were classified by the cross-validation method. In the classification using linear kernel function, the sensitivity of the positive sample was 90% and that of the negative sample was 91%. In the classification using radial basis kernel function, the sensitivity of the positive samples was 91%, and that of the negative samples was 90%. The purpose of this paper was to identify the thief zone, that is, to identify the negative sample. Therefore, the linear kernel function with the highest accuracy and the highest negative sample sensitivity were ultimately selected. Although many studies have reported that, in most cases, the radial basis kernel function has the strongest classification ability, it can be seen from the figure above that the linear kernel function has the highest accuracy and the strongest recognition ability with respect to the negative samples. This shows the necessity of the optimized kernel function and improves the reliability and stability of the model.
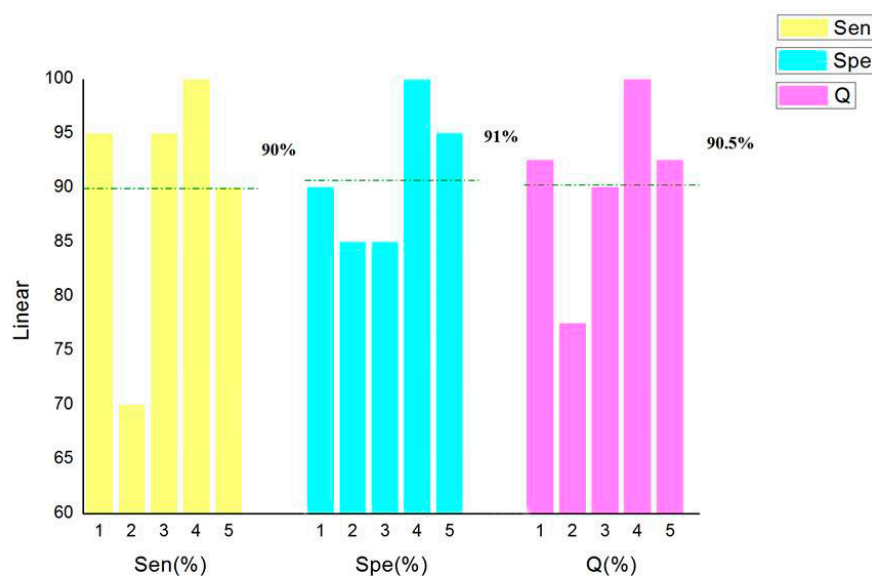

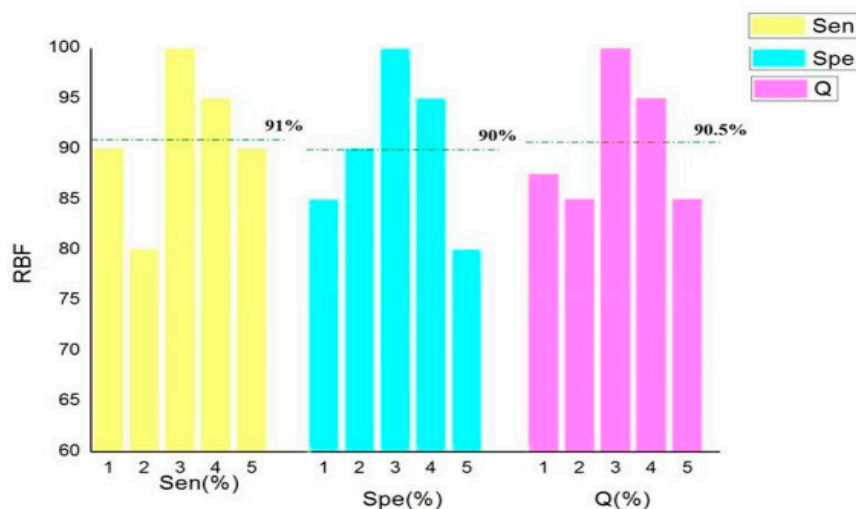**Figure 3.** Cross-validation results of linear kernel function.

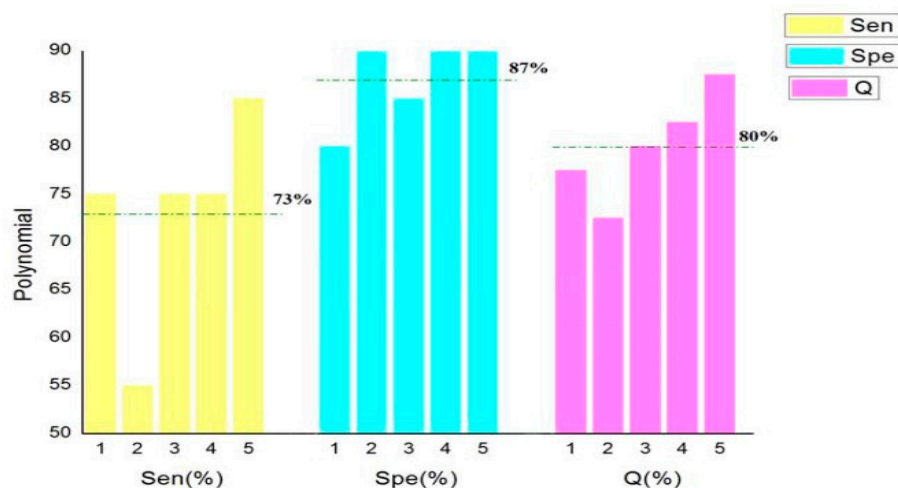**Figure 4.** Cross-validation results of radial basis kernel function.

Figure 5. Cross-validation results of polynomial kernel function.
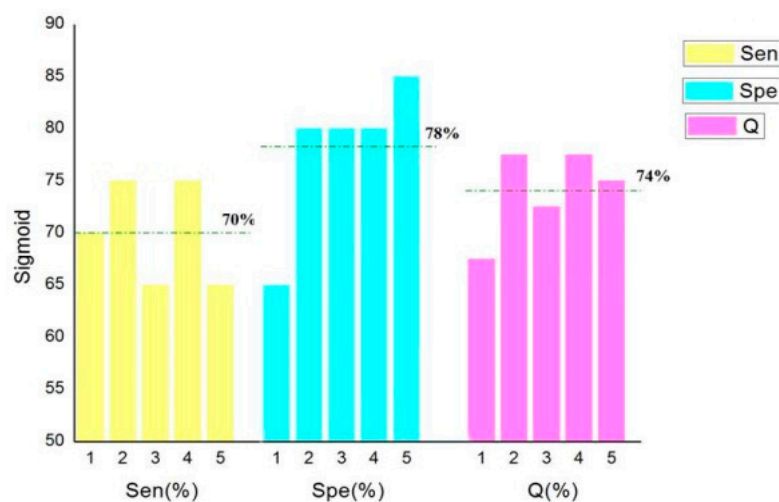

Figure 6. Cross-validation results of sigmoid kernel function.

The importance of each index of the thief zone identification model obtained by using the optimized kernel function (the linear kernel function) and the sample database to the classification is shown in Figure 7.

It can be seen in Figure 7 that porosity had the largest impact on classification, with a weight of 27%. This was because, after a long period of water flooding in M oilfield, the radius of the local pore throat in the reservoir increased, the degree of cementation weakened, and the porosity increased. Moreover, the oilfield has entered the later stage of development, and the water flooding development has failed to achieve the ideal oil-displacement effect. Therefore, a large amount of injected water rushed along the thief zone, resulting in an inefficient or even invalid circulation of injected water in many blocks. Porosity has become an important index affecting the formation of the thief zone. The second most important index was permeability, with a weight of 22%. This is because the greater the permeability, the smaller the vertical migration resistance of the water in the formation during water flooding development, causing the water to gradually converge at the bottom of the permeable layer along the formation direction, which forms the thief zone. The next most important indices were cumulative water injection, cumulative liquid production, and water cut, with weights of 17%, 15%, and 11%, respectively. Cumulative water injection, cumulative liquid production, and water cut were the dynamic response indices of the thief zone. Based on the comprehensive analysis of the cumulative water injection and the cumulative liquid production, when the cumulative water injection is constant, the larger the cumulative liquid production, the better the connectivity between the production well

and the water injection well, and the greater the possibility of forming the thief zone. The value of water cut reflects the amount of water in the produced fluid and the amount of water injected into the production well. The higher the water cut, the better the connectivity between the production well and the injection well, and the easier it is to form the thief zone.
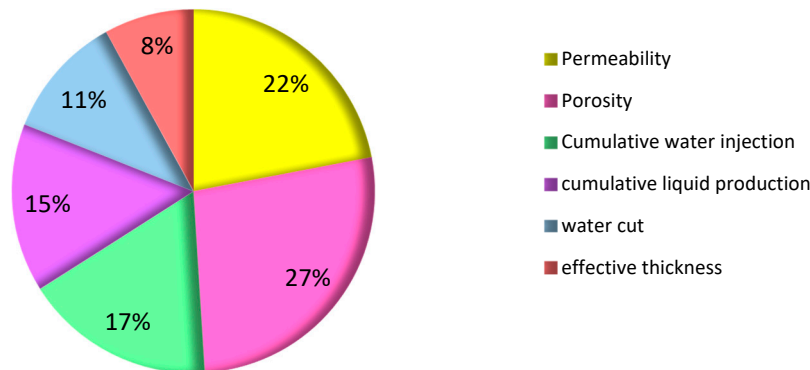

Figure 7. Importance of indices to classification.

The above model was applied to identify the thief zone of 82 groups of wells injected with tracer in M oilfield. There were four kinds of tracers used in the identification, namely NH_4SCN, I^{135}, NH_4NO_3, and $C_{28}H_{20}N_3O_5$. The earliest tracer injection occurred on 3 May 2014, and the latest occurred on 16 September 2016. The first tracer ended on 4 July 2016, and the last ended on 19 December 2017. The average breakthrough time of tracer was 6.4 months. The fastest breakthrough time was 2.1 months, and the slowest was 7.8 months. The identification result of the well group in which the tracer was detected was −1, while that of the group without the tracer detected was 1. The identification results and tracer identification results are shown in Figure 8. When the blue dot is inside the red square, then the algorithm correctly identified a thief zone, when it is not, it means wrong recognition.
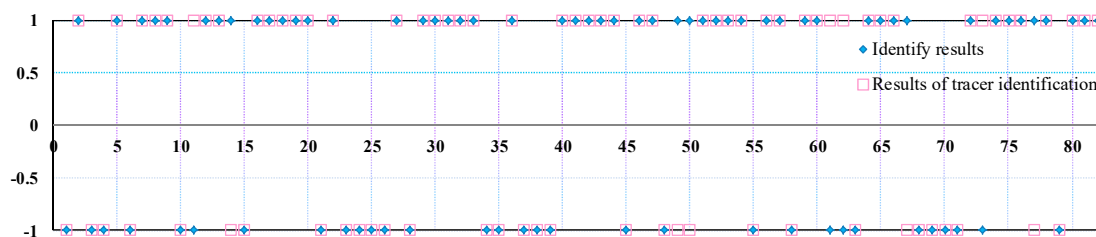

Figure 8. Comparison of identification results.

As can be seen in Figure 8, 27 groups of thief zone wells and 46 groups of non-thief zone wells were identified by the two methods. Additionally, 32 groups of thief zone wells and 50 groups of non-thief zone wells were identified by tracer. Thus, the recognition accuracy of the negative samples was 84.375%, that of the positive samples was 92%, and the total accuracy was 89.02%. It has been proven, therefore, that the method has a high total recognition accuracy and can provide certain reference value for the formulation of profile control and water plugging schemes in the high water cut period of an oil reservoir.

4. Conclusions

In this paper, a support vector machine was used to identify thief zones in a mature oil reservoir. The principle and identification flow of the method were introduced. Then, the method was applied to 82 groups of wells in M oilfield and verified by the tracer monitoring method.

(1) Through the analysis of the SVM identification results and tracer identification results, it was found that the thief zone of M oilfield is seriously developed, and 33% of the 82 well groups identified

by the two methods were identified as thief zones. This situation should be paid attention to and corresponding profile control and water shutoff measures should be taken to improve the thief zone.

(2) Through the analysis of the index weight affecting the thief zone, it was found that porosity has the largest influence on the formation of the thief zone, followed by permeability and then cumulative water injection, cumulative liquid production, and water cut.

(3) By using signal-to-noise ratio and correlation analysis methods to screen the indices, it was found that our proposed method can effectively avoid the problem of the classification accuracy of SVM being too low due to too many indices and the correlation between various indices.

(4) By comparing the SVM method with other reservoir engineering methods to identify the thief zone, it was found that the SVM method can effectively avoid the influence of human factors on the identification results, and the identification results were more objective and accurate.

(5) The identification results were verified by the tracer monitoring method, and it was found that the total accuracy of the SVM method in identifying the thief zone was 89.02%, that of the positive sample was 92%, and that of the negative sample was 84.375%.

Author Contributions: C.F. and T.G. put forward the idea of the identification in this paper and wrote the paper; B.H. designed the identification process; Y.W. and C.L. contributed to the results analysis. All authors reviewed the manuscript.

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

References

1. Robertson, J.O., Jr.; Oefelein, F.H. Plugging thief zones in water injection wells. *J. Pet. Technol.* **1967**, *19*, 999–1004. [CrossRef]
2. Sun, M.; Li, Z.P. Identification and description of preferential percolation path for waterflooding sandstone reservoir. *Fault Block Oil Gas Field* **2009**, *16*, 50–52.
3. Abdus, S.; Ghulam, M. Iqbal.16–Waterflooding and waterflood surveillance. In *Reservoir Engineering*; Abdus, S., Ed.; Gulf Professional Publishing: Amsterdam, The Netherlands, 2016; pp. 289–312.
4. Sayarpour, M. Field Applications of Capacitance-Resistance Models in Waterfloods. *SPE Reserv. Eval. Eng.* **2009**, *12*, 853–864. [CrossRef]
5. Albertoni, A.; Lake, L.W. Inferring Interwell Connectivity Only From Well-Rate Fluctuations in Waterfloods. *SPE Reserv. Eval. Eng.* **2003**, *6*, 6–16. [CrossRef]
6. Wang, X.; Xia, Z.J.; Zhang, H.W.; Liu, X.P.; Li, X.Q.; Zhang, L.J. Using Injection Profile Log Data to Distinguish Macropore Formation. *Well Logging Technol.* **2002**, *26*, 162–164.
7. Meng, F.S.; Sun, T.J.; Zhu, Y.; Feng, Q.F. A Study on the Method to Identify Large Pore Paths Using Conventional Well Logging Data in Sandstone Reservoirs. *Period. Ocean Univ. China* **2007**, *37*, 463–468.
8. Meng, F.S.; Huang, F.S.; Song, D.C.; Liu, G. Distinguishing Large Pore Paths in Sandstone Oil Layers by Fisher Method Using Logging Curves. *Period. Ocean Univ. China* **2007**, *37*, 121–124.
9. Al-Dhafeeri, A.M.; Nasr-El-Din, H.A. Characteristics of high-permeability zones using core analysis, and production logging data. *Pet. Sci. Eng.* **2007**, *55*, 18–36. [CrossRef]
10. Li, G.J.; Liang, J.; Li, W. A Study on the Method to Identify Large Pore Paths Using Well Logging Data. *Oil Gas Field Surf. Eng.* **2008**, *9*, 11–12.
11. Li, B.J.; Hamad, N.; Jim, L.; Mansoor, A.R.; Ihsan, G.; Mohammed, A.K. Detecting thief zones in carbonate reservoirs by integrating borehole images with dynamic measurements. In Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, CO, USA, 21–24 September 2008.
12. Chen, Q.; Gerritsen, M.Q.; Kovscek, A.R. Effects of reservoir heterogeneities on the steam assisted gravity drainage process. *SPE Reserv. Eval. Eng.* **2008**, *11*, 921–932. [CrossRef]
13. John, D.; Hans, V.D.; Maersk, O.; Arve, O.N. Interwell communication as a means to detect a thief zone using DTS in a Danish Offshore well. In Proceedings of the SPE Offshore Technology Conference, Houston, TX, USA, 6–9 May 2013.

14. He, C.Z.; Hua, M.Q. Fractal Geometry Description of Reservoir Pore Structure. *Oil Gas Geol.* **1998**, *1*, 17–25.
15. Liao, M.G.; Li, S.L.; Tan, D.H. Relationship Between Permeability and Mercury Injection Parameters Curve for Sandstone Reservoir. *J. Southwest Pet. Inst.* **2001**, *4*, 5–8.
16. He, Y.; Yang, C.M.; Ying, J.; Yan, J.H. A New Method for Measuring the Quantitative Pore Throat Volume. *J. Southwest Pet. Inst.* **2002**, *3*, 5–7.
17. Shi, Y.G.; Zeng, Q.H.; Zhou, X.J. Interpreting Model of Large Pore well Testing Theory. *Oil Drill. Prod. Technol.* **2003**, *25*, 48–50.
18. Yin, W.J.; Chen, Y.S.; Wang, H.; Zhang, J.T.; Zhang, X.M.; Ding, Y.M. Building of the Interpretation Model of the Large Channels and Remaining Oil Saturation by Hydraulic Survey. *Pet. Geol. Recovery Effic.* **2005**, *12*, 63–65.
19. Feng, Q.H.; Wang, S.; Zhang, W.; Song, Y.; Song, S. Characterization of high-permeability streak in mature waterflooding reservoirs using pressure transient analysis. *Pet. Sci. Eng.* **2013**, *110*, 55–65. [CrossRef]
20. Liu, H. Well Testing Interpretation Method Research of Preferential Seepage Channels. *Pet. Geol. Eng.* **2015**, *29*, 98–100.
21. Izgec, B.; Kabir, S. Identification and characterization of high-conductive layers in waterfloods. *SPE Reserv. Eval. Eng.* **2009**, *14*, 113–119. [CrossRef]
22. Batycky, R.P.; Thiele, M.R.; Baker, R.O.; Chung, S. Revisiting Reservoir Flood-Surveillance Methods Using Streamlines. *SPE Reserv. Eval. Eng.* **2008**, *11*, 387–394. [CrossRef]
23. Wang, Y.Q.; Chen, F.H.; Gu, H.J.; Zhou, H.Z.; Nie, Z.R.; Liu, F.K.; Lu, M.H. Using Tracer to Study Interwell Water Flow Predominant Channel. *Xinjiang Pet. Geol.* **2011**, *32*, 512–514.
24. Dou, Z.L.; Zeng, L.F.; Zhang, Z.H.; Xiong, W.; Tian, G.L.; Liu, X.W.; Huang, L.X. Research on the diagnosis and description of wormhole. *Pet. Explor. Dev.* **2001**, *28*, 75–77.
25. Zeng, L.F.; Chen, B.P.; Wang, X.Z. Preliminary study on quantitative description of high capacity channel in loose sandstone reservoir. *Pet. Geol. Recovery Effic.* **2002**, *9*, 53–54.
26. Liu, Y.T.; Sun, B.L.; Yu, Y.S. Fuzzy Identification and Quantative Calculation Method for Big Pore throat. *Oil Drill. Prod. Technol.* **2003**, *54*, 95–96.
27. Wang, S.L.; Jiang, H.Q. Determine level of thief zone using fuzzy ISODATA clustering method. *Transp. Porous Media* **2010**, *86*, 483–490. [CrossRef]
28. Udy, J.; Hansen, B.; Maddux, S.; Petersen, D.; Heilner, S.; Stevens, K.; Hedengren, J. Review of Field Development Optimization of Waterflooding, EOR, and Well Placement Focusing on History Matching and Optimization Algorithms. *Processes* **2017**, *5*, 34.
29. Ding, S.W.; Jiang, H.Q. Identification and characterization of high-permeability zones in waterflooding reservoirs with an ensemble of methodologies. In Proceedings of the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Nusa Dua, Indonesia, 20–22 October 2015.
30. Huang, B.; Xu, R.; Fu, C.; Wang, Y.; Wang, L. Thief Zone Assessment in Sandstone Reservoirs Based on Multi-Layer Weighted Principal Component Analysis. *Energies* **2018**, *11*, 1274. [CrossRef]
31. Huang, B.; Xu, R.; Fu, C.; Zhang, W.; Shi, Z.Z. Research on multi-level fuzzy comprehensive identification method for the interwell thief zone in sandstone reservoir. *Lithol. Reserv.* **2018**, *30*, 105–112.
32. Kamari, A.; Bahadori, A.; Mohammadi, A.H. An Efficient Approach for the Determination of Oil Production Rate During the Water-flooding Recovery Method. *Pet. Sci. Technol.* **2015**, *33*, 1208–1214. [CrossRef]
33. Jin, Z.Y. Application of SVM in Identification of High Permeability Channels. *Pet. Geol. Eng.* **2009**, *28*, 178–180.
34. Chen, C.L.; Wang, Z.; Niu, W.; Wang, H.M.; Guo, L.F. Quantitative calculation method of thief zone based on least square support vector machine. *Fault Block Oil Gas Field* **2015**, *22*, 74–77.
35. Salcedo-Sanz, S.; Rojo-lvarez, J.L.; Marnez-Ram, M. Support vector machines in engineering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4*, 234–267. [CrossRef]
36. Li, K.; Li, C.L.; Zhang, W. Research of Diver Sonar Image Recognition Based on Support Vector Machine. *Adv. Mater. Res.* **2013**, *785*, 1437–1440. [CrossRef]
37. Vanhoy, G.; Schucker, T.; Bose, T. Classification of LPI radar signals using spectral correlation and support vector machines. *Analog Integr. Circuits Signal Process.* **2017**, *9*, 305–313. [CrossRef]

38.   Taş, E. Classification of Gene Samples Using Pair-Wise Support Vector Machines. *Alphanumer. J.* **2017**, *5*, 283–292.

39.   Wang, H.F.; Zheng, B.C.; Yoon, S.W.; Ko, H.S. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Oper. Res.* **2018**, *267*, 687–699. [CrossRef]