This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Collusion Strategy Investigation and Detection for Generation Units in Electricity Market Using Supervised Learning Paradigm

Peyman Razmi, Majid Oloomi Buygi, *Member, IEEE*, and Mohammad Esmalifalak, *Member, IEEE*

***Abstract*—In an oligopoly market, producers compete together to seize the electricity market share. Since they cannot obtain their desired profits through fair competition, they may collude to set their bid prices illegally higher than the oligopoly level. Manipulation and increasing market price decrease social welfare and then market efficiency. This article intends to provide independent system operators (ISOs) with a tool to analyze day-ahead market data so as to identify generator units who intend to exercise collusion and raise the market prices. Toward this goal, all possible collusion and competition scenarios are simulated and then the generated data are used to train a supervised learning algorithm. By applying the proposed approach to the IEEE 57 and 30 bus test systems, the efficiency of the proposed approach was assessed. Furthermore, it is demonstrated how colluding generators choose between maximizing their colluded profit and reducing the risk of being detected by ISO. The results show machine learning is capable of identifying colluding companies with accuracy of 95%. Also, it was rightly obvious that the closer the bidding price of companies is to competitive level, the more downward the efficiency of the machine is.**

***Index Terms*—Collusion, equilibrium point, machine learning algorithm, nonoptimal strategy.**

I. Introduction

In a competitive electricity market, producers cannot affect the market price; in other words, they take the price from the market and attempt to determine the amount of their own production according to the market price. Therefore, the optimum production level will be calculated by the intersection of the marginal cost curve with market price. Limitations of power systems have pushed the electricity markets from competitive environment toward oligopoly market. In an oligopoly market, there is a possibility for producers to affect the market price. Producers, who are interested in raising the price of the market, may increase their bids (economic withholding) or reduce the production (physical withholding) to influence the market price in their favorable direction. Producers who can affect the market with any of the aforementioned methods have so-called "market power" [1], [2]. In addition to raising prices higher than the competitive level, the ability of maintaining market price for a long period of time is important. After liberalization of the electricity market, price manipulation and market power of individual producers [3]–[5] can be resulted from the joint decision and alliance of two or several producers that in such cases, collusion is done in the form of implicit and explicit [6], [7]. A long-term alliance between several companies to inflate the market price out of its competitive range is called collusion. Generation companies, which cannot obtain their desired profit through fair competition in oligopoly market, may collude and set their bid prices and consequently direct market price to a value abnormally higher than what oligopoly competition commands. Explicit collusion is a hidden agreement for interaction among electricity power producers who share confidential information in order to control the market price. The goal of a coalition is to adopt or enforce unified strategies to increase the profit of its members. However, influence of a coalition does not remain limited to the bids of its members and will alter the bids of its competitors. Stakeholder's choice of coalition is based on the projected profit after the formation of coalition. In other words, each agent selects the coalition that provides maximum profit. Members of an explicit coalition share the profit that is obtained through price manipulation [8], [9]. With the advent of modern wholesale electricity markets in developed countries, these markets have become the subject of many discussions. These discussions often originate from the nature of these electricity markets, e.g., the facts that they are often controlled by a few numbers of companies, the traded commodity (electricity) cannot be stored, demand has an inelastic nature, and technical limitations, such as congestion of transmission line, can lead to isolation of submarkets. These characteristics provide an enabling environment for coalitions, which can be regarded as collusion [10]. The absence of competition in a market where prices are set by collusion leads to not only the violation of consumers' rights, but also the reduction of producer's efficiency. In light of these adverse effects, regulatory organizations often prohibit such coalitions to protect competition [11], [12]. Therefore, in order to maintain the competition in the electricity market and reduce exercising of collusive behavior from producers and improve market efficiency, this article intends to provide independent system operators (ISOs) with a tool to analyze day-ahead market data so as to identify generator who intend to exercise collusion and raise market prices. This article focuses on the explicit collusion

Manuscript received September 1, 2019; revised March 23, 2020; accepted April 22, 2020. *(Corresponding author: Mohammad Esmalifalak.)*

Peyman Razmi and Majid Oloomi Buygi are with the Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad 108847, Iran (e-mail: peyman.razmi65@gmail.com; m_oloomi@yahoo.com).

Mohammad Esmalifalak is with the McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: esmalifalak.mohammad@gmail.com).

Digital Object Identifier 10.1109/JSYST.2020.2991608

1937-9234 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

and proposes an approach to reveal and detect such collusive behaviors. Automatic detection of such collusive behaviors is not an easy task. The basic approach is to use supervised learning, but the major problem of this approach is the unavailability of data specifically referring to collusion. The alternative is to use unsupervised learning techniques for this purpose [13], [14]. The detection mechanism introduced in [13] focuses on circular trading, in which a ring of colluders trades a certain share repeatedly to raise their price. The proposed method is based on interpretation of stock flow graph with a researcher who developed Markov clustering algorithm. Palshikar and Apte [14] focused on the detection of cross trading and investigated the aptitude of different clustering algorithms for collusion detection in electricity markets. Mihailescu and Ossowski [15] detected collusion using a two-step process in an energy market. First, the behavioral pattern of collusion is investigated and then using a change-point analysis, the behavior of each agent is studied to reveal the possible structural breakpoints. Existence of such breakpoints could be a sign for collusion and has to be analyzed more in a verification phase, where behavioral similarities of candidates are checked using statistical methods. As an instance, market power execution and inefficiencies of Alberta's restructured electricity market are studied in [4] using hourly wholesale market dataset from 2008 to 2014. Authors found that firms conduct considerable market power in the highest demand hours with bounded excess generation capacity. Aliabadi *et al.* [16] used a game-theoretic model to analyze the behaviors of ISO and generator and the situations leading to collusive transactions by generators. Zideh and Mohtavipour [17] provided an approach to analysis of the development of tacit collusion between a Genco (generator company) and a Disco (distribution company) in a simulated electricity market. Authors modeled Gencos' and Discos' behaviors using the 'SARSA learning algorithm and a model was used to tune continual exploration and make a tradeoff between exploration and exploitation.

This article provides an approach for ISOs in which day-ahead market data are analyzed so as to identify generator units that collude together and increase the market price illegally. High cost of labeling datasets by domain experts has made the labeled data a rare find in real-world applications. While one approach is to use unsupervised algorithms, our goal is to produce synthetic data to train and validate our algorithm. We have not found labeled dataset in the literature (labeled neither by experts nor synthetic data created by researchers) and existing works have focused more on the potential of tacit collusion and its impact on the market. The main contribution of this article is two folds. First, it defines how we can create such synthetic data reliably, which could be a benchmark for future research on modeling new vulnerabilities. Second, contribution is to use clustering methods to detect colluding market participants. Toward this goal, this article uses Nash equilibrium theory and supervised learning methods. First, by using Nash equilibrium theory, market equilibrium and then historical data/quasi-actual data are computed for different load levels and for different collusion scenarios. In the second part, a machine learning technique is developed by using supervised learning paradigm so that market equilibrium data and their peripheral operating points as a quasi-actual market data are used to train the collusion-detecting machine. The rest of this article is organized as follows.

Section II describes the process through which Nash equilibrium is modeled and market equilibrium is computed. Section III overviews the machine learning methods and supervised algorithms that are proposed and used in this article to detect collusion. In Section IV, the approach of finding quasi-actual system operating points and collusion criteria are described, then the framework of data collection to train the machine learning is explained. In Section V, the proposed collusion detection framework is applied to the model of typical electricity market, and then, obtained results are discussed and analyzed in Section VI. Finally, Section VII concludes this article.

II. Modeling of Equilibrium Point

In practice, there is not enough data from different colluded scenarios to train the collusion-detecting machine. The purpose of equilibrium point modeling is to simulate and collect data for the training process in subsequent steps. Normal operating points are usually fluctuating around Nash equilibrium. Nash equilibrium is the point where no firm is better off by changing its strategy unilaterally [18]. In other words, deviation from the Nash equilibrium point by participants will not increase their profits. For the modeling of equilibrium point, it is necessary to define the assumptions and conditions of market operation in the restructured power industry. For this purpose, we assume that the day-ahead market is a pool-based market with uniform pricing. Assume the cost of generating Qs_i by unit *i* is as following:

$$C(Qs_i) = a_i Qs_i + \frac{1}{2} b_i Qs_i^2. \tag{1}$$

These coefficients reflect the operation cost of unit *i* when it generates Qs_i. Also, a_i and b_i are the cost function coefficients of unit i. Also, the utility of consuming Q_{Dj} by consumer j is

$$C(Q_{Dj}) = c_j Q_{Dj} - \frac{1}{2} d_j Q_{Dj}^2. \tag{2}$$

These coefficients reflect the utility of consumer *j* when he or she consumes Q_{Dj}. Moreover, c_j and d_j are demand function coefficients of consumer *j*. The true linear supply function of generation unit *i*, which is real marginal cost of unit *i*, is as follow:

$$\rho_{\text{true}}(Qs_i) = a_i + b_i Qs_i \tag{3}$$

where ρ represents price or the marginal cost of unit *i*. These marginal costs are confidential and strategic data for the producers. In the supply side, a set of companies is defined as following:

$$F = \{f1, f2, f3, \ldots, f_k\}. \tag{4}$$

The goal of firm *f*, the owner of unit *i* where $f \in F$, is to maximize its profit by determining the optimal parameters of a similar linear bid function as following:

$$\rho_{\text{bid}}(Qs_i) = \alpha_i + \beta_i Qs_i. \tag{5}$$

These coefficients reflect the purchasing cost from unit *i* when it generates Qs_i. To obtain a unique solution, one must avoid changing both α and β, and instead keep one parameter constant

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

(e.g., $\beta_i = b_i$) and alter the other parameter (α_i) [19]. The ISO supervises the schedule of generators and market clearing price (MCP) with the aim of maximizing social welfare without violating the technical constraints. The objective of ISO can therefore be modeled as following:

$$\text{Max} \quad J_{\text{ISO}} = \sum_{j \in D} \left(c_j Q_{Dj} - \frac{1}{2} d_j Q_{Dj}^2 \right) - \sum_{i \in S} \left(a_i Q s_i + \frac{1}{2} b_i Q s_i^2 \right) \tag{6}$$

s.t.

$$\sum_{i \in S} Q s_i - \sum_{i \in D} Q_{Di} = 0 \tag{7}$$

$$Q s_i^{\min} \leq Q s_i \leq Q s_i^{\max} \tag{8}$$

where J_{ISO} is the social welfare, *S* is the set of generation units, *D* is the set of consumers, and $Qs_i^{\min}$ and $Qs_i^{\max}$ are the capacity limits of unit *i*. Firm *f* maximizes its profit by determining the optimal bid for its units using the following bilevel optimization:

$$\text{Max} \quad \pi_f = \sum_{i \in f} \left(\lambda Q s_i - a_i Q s_i - \frac{1}{2} Q s_i^2 \right) \tag{9}$$

s.t.

$$\alpha_i^{\min} \leq \alpha_i \leq \alpha_i^{\max} \tag{10}$$

$$\text{Optimization problem of ISO (6)–(8)} \tag{11}$$

where π_f is the profit of firm *f*, S_f is the set of generation units of firm *f*, $\alpha_i^{\min}$ and $\alpha_i^{\max}$ are lower and upper limits of α_i, respectively, and λ is the power balance constraint (7). The strategy of firm *f* to outbid other firms is obtained by solving a mathematical program with equilibrium constraints (MPECs) that consists of (9)–(11). Decision variable of generating firm *i* is α_i. Market equilibrium is obtained by solving equilibrium problem with equilibrium constraints (EPECs) that consists the set of MPECs of all generating firms. To solve the EPEC, first constraint (11) is replaced by Karush–Kuhn–Tucker (KKT) optimality conditions of optimization (6)–(8). Then, KKT conditions for optimization problem of each generating firm are written. Finally, KKT conditions of all generating firms are solved simultaneously using dual variables based algorithm [19]. Dual variables based algorithm is an iterative algorithm. In each iteration, first active constraint with the biggest dual variable is identified and KKT conditions are revised. The algorithm is continued until all active constraints at market equilibrium are identified and KKT conditions are revised based on active constraints. The remaining KKT conditions are linear equations with unique market equilibrium (for more details, refer to the work in [19]).

III. Machine Learning

Machine learning can be described as the process through which a computer can learn to perform a task with new data or configurations without any reprogramming. The concept of machine learning originates from pattern recognition and computational learning theory in artificial intelligence [20]. The algorithms developed with machine learning concepts are expected to learn to do their tasks for future data, which is not presented to the algorithm during training process [21]. Machine learning has a multitude of tasks, the most important of which is perhaps the supervised learning [22]. In supervised learning, we provide the algorithm with certain inputs and a group of labeled outputs (targets), a certain output in the training process. Algorithm then uses machine inference to develop a function capable of emulating the process and mapping the new inputs data to the predicted output. In the following section, we have explained some of supervised machine learning algorithms that we have used in this article for collusion detection in electricity market.

A. Support Vector Machine (SVM)

In this section, SVM is proposed for detecting collusion and identifying the companies who have raised the market price illegally. In 1995, Vapnik and Cortes introduced SVM theory in which a hyperplane or a series of hyperplanes was constructed and utilized for both classification and regression [23], [24]. By considering the labeled training set *S* = (x_l, y_l), l = 1,. . . ,*L* of size *L*, and $y_l \in$ [1,−1]. The SVM can be obtained by solving

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + \sum_{i=l}^{L} \xi_l \tag{12}$$

s.t.

$$y_i (w^T \phi(x_l) + b) \geq 1 - \xi_l \tag{13}$$

$$\xi_l \geq 0, l = 1, \ldots, L \tag{14}$$

where $\phi(x_l)$ represents a nonlinear transformation mapping x_l in a high-dimensional space that is called kernel function. The slack variable ξ_l represents nonlinearly separable training sets, and C denotes the parameter of a tunable positive regularization. In order to achieve a distributed SVM, (12) can be rewritten as follows:

$$\min_{w_i, b_i, \xi_i} \quad \frac{1}{2} \sum_{i=1}^{N} w_i^T w_i + C \sum_{i=1}^{N} \sum_{l=l}^{L} \xi_{il} \tag{15}$$

s.t.

$$y_{il} (w_i^T \phi(x_{il}) + b_i) \geq 1 - \xi_{il} \tag{16}$$

$$\xi_{il} \geq 0, i = 1, \ldots, N, l = 1, \ldots, L \tag{17}$$

where N denotes the number of groups working together in order to train the SVM and w_i represents the parameter of the local optimization for each group. By introducing a global variable z, (15) can be reformulated as

$$\min_{z, w_i, b_i, \xi_i} \quad \frac{1}{2} \sum_{i=1}^{N} z^T z_i + C \sum_{i=1}^{N} \sum_{l=l}^{L} \xi_{il} \tag{18}$$

$$y_{il} (w_i^T \phi(x_{il}) + b_i) \geq 1 - \xi_{il} \tag{19}$$

$$\xi_{il} \geq 0, z = w_i \tag{20}$$

$$i = 1, \ldots, N, l = 1, \ldots, L. \tag{21}$$

In order to solve (18) distributively, variables $\{z, w_i\}$ i =1,..., *N* can be partitioned into two sets represented by $\{z\}$ and $\{w_i\}$,

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$i = 1$,...,N, and the alternating direction method of multipliers can be applied to solve the problem. Specifically, the scaled augmented Lagrangian function can be expressed as follows:

$$l\{z, w_i, \xi_i, \rho, \mu_i\} = \frac{1}{2}\sum_{i=1}^{N} z^T z_i + C\sum_{i=1}^{N}\sum_{l=l}^{L}\xi_{il} + \frac{\rho}{2}\|w_i - z + \mu_i\|_2^2 \tag{22}$$

where ρ denotes the step size and μ_i represents the scaled dual variable. At each iteration *k*, $\{w_i\}$,{z}, and μ_i can be updated as follows:

$$w_i[k+1] = \arg\min_{w_i, b_i, \xi_i} C\sum_{i=1}^{L}\xi_{il} + \frac{\rho}{2}\|w_i - z[k] + \mu_i[k]\|_2^2 \tag{23}$$

$$y_{il}(w_i^T \phi(x_{il}) + b_i) \geq 1 - \xi_{il} \tag{24}$$

$$\xi_{il} \geq 0, l = 1, \ldots, L. \tag{25}$$

Note that the process updating w_i can be done locally in the ith group. Moreover, it involves the fitting of an SVM to the local data using an offset in the quadratic regularization term. The vector {z} is expressed as

$$k_i[k+1] = \underset{z}{\text{argmin}}\frac{1}{2}z^T z + \frac{\rho}{2}\|w_i[k+1] - z + \mu_i[k]\|_2^2 \tag{26}$$

which can be solved analytically as

$$z = \frac{N_\rho}{\frac{1}{\rho} + N_\rho}(\bar{w}[k+1] + \bar{\mu}[k]) \tag{27}$$

in which $\bar{w} = \frac{1}{N}\sum_{i=1}^{N} w_i$ and $\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mu_i$. Finally, the scaled dual variable μ_i can be updated by

$$\mu_i[k+1] = \mu_i[k] + w_i[k+1] - z[k+1]. \tag{28}$$

In constructing historical data, since the number of classes and classes in collusion samples depend on the number of generators and production companies as well as the number of possible collusion scenarios, so the number of classes can be more than two groups, therefore, a multiclass SVM is used [25], [26], and for each nonlinear classifier, Gaussian kernel is applied in which δ is a mutable parameter. The main factors affecting the performance of SVM comprise kernel function and its parameters as well as the soft margin parameter C. The optimal Gaussian kernel parameter and soft margin (*C*) can be used for improving the efficiency of nonlinear SVM. Gaussian kernel, which has a single parameter (γ), is a typical choice for SVM [27]. The common practice for finding the best values of C and γ is to conduct a grid search, i.e., to repeat the calculations with different C and γ combinations and determine the values yielding the best accuracy through cross-validation [28]. The SVM algorithm is mentioned in [29].

B. Classification and Regression Tree (CART)

Briman *et al.* introduced CART algorithm, resulting in the creation of a tree using binary division [30], [31]. CART algorithm is considered as a nonparametric decision tree learning technique, which can be used to classify various types of data. In other words, both CARTs are generated by the CART algorithm, which can be dependent on whether the dependent variable is categorical or numerical, respectively. CART algorithm is a classification technique for building a decision tree based on Gini's impurity index as splitting criterion. CART is a binary tree that is constructed by dividing the node into two child nodes repeatedly. Algorithm 3 describes the process of constructing a CART algorithm, as we see it works repeatedly in three steps based on Gini's impurity index. In the first step for each feature, the best split is calculated, which maximizes the splitting criterion. In step 2, the node's best split is selected among the best splits from step 1. In step 3, CART algorithm splits the node by using the best node split from step 2 and repeats from step 1 until stopping criterion is satisfied (see Appendix B and [32]).

C. Bootstrap Aggregating Method (Bagging)

Bootstrap aggregating, also known as bagging, is designed as a machine learning ensemble meta-algorithm for the improvement of both the stability and accuracy of machine learning algorithms employed in the statistical regression and classification. Moreover, it can decrease the variance and prevent overfitting. Breiman proposed bagging for the improvement of the classification via combining classifications of randomly produced training sets. Bagging method is modeled based on the instability of base learners that can be utilized to modify the predictive performance of such unstable base learners. The main idea is that there is a training set S of size n and a learner L, which commonly is decision tree, bagging create m new training sets with replacement S_i. Then, bagging applies L to each S_i to build m models. The final output of bagging is based on simple averaging (see bagging algorithm in Appendix B) [33].

D. Statistical Anomaly Detection Techniques

In data mining, the datasets that are considerably different from the remainder of the data are called outliers or anomalies. Different types of anomaly detection methods have been proposed, such as the distance-based, model-based, and statistical-based methods [34]. In this article, we use the statistical-based methods. We use metric $P(z)$ and a threshold δ, where $P(z)$ represents the statistical characteristics of the historical data. If $P(z) \leq \delta$, then z statistically has low similarity to the remaining data. In this method, the hypothesis of anomaly is confirmed if $P(z) \leq \delta$, and it is rejected if $P(z) \geq \delta$. Threshold δ will be learnt by the historical data (step 3 in Algorithm 4). Because of using the historical data to learn δ, this method, in some literature, is called the semisupervised learning method. We use the multivariate Gaussian distribution probability density function (pdf) as metric $P(z)$ as follows:

$$P\left(Z; \mu, \sum\right) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{0.5}} \times \exp\left[-\frac{1}{2}(Z-\mu)^T \sum^{-1}(Z-\mu)\right]$$

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$$\mu = \frac{1}{m}\sum_{i=1}^{m} Z^{(i)}$$

$$\sum = \frac{1}{m}\sum_{i=1}^{m}\left(Z^{(i)} - \mu\right)\left(Z^{(i)} - \mu\right)^{T} \qquad (29)$$

where n is the number of features, m is the number of samples, and $P_i(z_i)$ is the pdf of feature i. Each feature z_i follows a certain distribution that should be fitted based on the historical data. In step 1, the historical data will be collected. These data will be gathered from the ISO. In step 2, a Gaussian density function will be fitted to the preprocessed data $(P(Z))$. If the new operational points are statistically not similar to the historical data, the value of $P(Z\text{val})$ will be less than threshold δ. Step 3 defines the best possible δ using the labeled historical data. The new operating point will be tested in step 4 to see how similar it is to the normal data.

IV. Collusion Detection and Classification

This article intends to provide ISOs with a tool to analyze day-ahead market data so as to identify generator units who intend to exercise collusion and raise the market prices. This objective is pursued via machine learning and supervised approaches. One major problem that makes it unable to use trainable machines in collusion detection area is lack of adequate data associated with different collusion among generation companies. To address this issue, we first simulate the electricity oligopoly market and then create quasi-actual operating points for different collusion scenarios. To this end, we have to study scenarios which have the highest impact on electricity market and threat the competitive environment of power market. These collusion scenarios are well studied in research community, as an example in [8] and [9], authors have modeled the collusive strategic behaviors, including implicit or explicit collusion. To create the quasi-actual operating points for collusion strategies, we need to model electricity market for different collusion scenarios. In order to model collusion scenarios, the companies participating in coalition among themselves are considered as a single entity. This means that generating firms (Gencos) that participate in the collusion seeks optimization of the sum of profits of all participated firms in the collusion, instead of optimizing their profits separately and independently. In order to create the historical data about all different collusion scenario, the quasi-actual operating points of different bilateral collusion are calculated. At first, our article focuses on identifying specific collusion and collusion-free scenarios and labeling the data as the operating points of the market, then the data are used to train supervised algorithms. In other words, our goal is to use supervised algorithms to detect the specified collusion scenarios. In this article, we present a practical approach for generating electricity market operating points and day-ahead market data as training and evaluating data for detecting machine.

A. Data Generation

In the generation of training and test data process, we need two points of view. The first view comes from the ISO to create training data and developed the trainable machines. The second view point is Gencos's view to create test and evaluation data in order to assess the trained machines. Below we will discuss how to generate training and evaluating data from two viewpoints.

1) Generation of Training Data by ISO's View Point: As mentioned in previous section, in order to collusion detection and identifying the colluding companies, system operator needs to create and design a trainable machine. To do this, we have to simulate the quasi-actual operating points and market data from ISO's viewpoint based on Nash equilibrium theory for competitive or collusion-free state and different collusion scenarios according to Section II. In this article, some uncertainties in cost function of generation units are added in the generating of quasi-actual data process systematically as a more practical approach, therefore, it is assumed that system operator is unable to access adequate knowledge about some cost function of generator. Most importantly, in this article, in order to propose and provide a practical approach, training data are created by system operator to model a trained machine and test data are generated from Gencos's view to assess created model, which will be described in the next section. In the generation of training data process, we generate market equilibrium points from ISO's viewpoint, which are intercept of bids of all generating units (α). Training data are generated by these equilibrium points. To this end, first we solve EPEC problem for many times by using Monte Carlo theory that estimates cost function of all generator. Therefore, from ISO's viewpoint by using Monte Carlo theory, according to (3) for cost function coefficients a_i, we consider a set of possible coefficients for each generator, for example, for generator G_1, we have $[a_{11}\ a_{12}\ a_{13}\ ...a_{1n}]$ where n is the number of cost function scenarios for each generator and assuming that coefficient b_i is known for all sides, then the EPEC problem is solved for many times and different scenarios of all possible generator's cost function coefficients. Thus, too many equilibrium points are created for different collusion scenarios. To achieve a proper collusion detection technique, modeling should go beyond the limits of a single period or a certain class of load demand, therefore generation of training data is repeated for different load levels. Since for each generator, equilibrium points are obtained for different possible cost function coefficients, only one of them is considered as actual market equilibrium points and the rest are close to equilibrium's peripheral points, therefore obtained equilibrium points and their peripheral points are assumed as the main core of quasi-actual operating points in the training process. It should be noted that the obtained equilibrium point is not enough to create market data or quasi-actual operation point and does not contain useful information about collusive behavior of Gencoes to construct a model of trained machine for collusion detection, therefore, in addition to equilibrium points, some attributes are required to model of detector machine and distinguish operating points related to collusion from the collusion-free operating points. Using these features, we can find out the collusive behaviors of colluding companies. These attributes are utilized as inputs to train and model the trainable machine. These attributes are as follows:

1) marginal cost of generators (MC);
2) MCP of generators;

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1: Training Algorithm.

```
1 Historical data generation
$a$)Consider a limit for marginal costs
$b$)Select different scenarios for marginal costs
$c$)Consider different scenarios for collusion state
for $SC = 1 : SC_{max}$ (collusion scenarios) do
    for $S_{MC} = 1 : k$ (marginal cost scenarios) do
        for $LD = Q_{Di} : \Delta Q_D : Q_{Dmax}$ do
            $a$)Compute market equilibrium
            $b$)Compute generation powers at
              equilibrium
            $c$)Calculate collusion criteria
            x←[MC, MCP, Ler, Market Share, HHI]
        end
        Save criteria and label for each scenario
        X← x
        Y← sc   (label)
    end
    Data ←[X Y]
end
2 Train ML algorithm[22-23]
```

3) Lerner index for different generators (Ler);
4) market share of each producer in each load demand; and
5) market Herfindahl Hirschman Index (HHI).

Privately owned generators naturally refuses to publicly report their marginal cost functions, therefore in this section, it is assumed that ISO is unaware about marginal cost functions of some generators. The Lerner index measures the extent to which a given firm's price exceeds its marginal costs. In other words, the Lerner index measures a firm's level of market power by relating price to marginal cost and describes the relationship between elasticity and price margins for a profit-maximizing firm. The HHI index is famous and accepted as an indicator of market competition that measures the level of concentration in a given industry [27]. The market simulation performed in this article aims to calculate the aforementioned criteria at equilibrium points and their peripheral points that are under different collusion and collusion-free scenarios. Now, we were able to create training and historical data and model a trained machine with the help of Nash theory. After the training process, new samples of the real market are given to the trained machine and the machine detects the occurrence of collusion and distinguishes companies who participate in the collusion. Algorithm 2 shows the basic procedure of the historical and training data generation from ISO's viewpoint. In step 1, historical data will be collected. First, we consider a limit for marginal costs of all generators (parameter a) from ISO's viewpoint, then consider different values for parameter a for each generator (according to Monte Carlo theory). After that we consider all possible combinations and scenarios for marginal costs of generators separately in modeling of game theory between companies and modeling of the collusion and collusion-free states. The first and second loops are related to collusion and marginal cost scenarios, respectively, and the third loop is related to different load levels. Therefore, in each collusion scenario, we compute market equilibrium, generation powers at equilibrium point, and then collusion criteria for all marginal cost scenarios and all load levels. The obtained data in each collusion scenario are labeled and saved, finally data generation process is repeated for other collusion scenarios. In the last step, the obtained collusion criteria are utilized as input to train the detector machine.

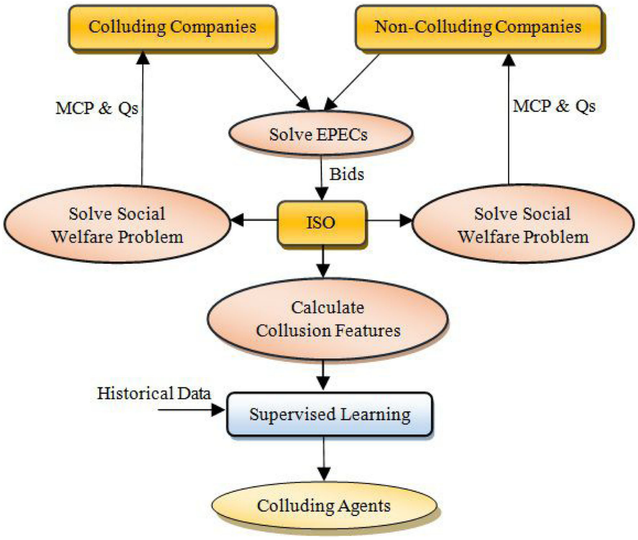

Fig. 1. Schematic of the proposed method for collusion detection.

2) Generation of Test Data From Gencos's View: In previous section, we were able to model a trained machine via quasi-actual operating points, which were generated from ISO's view. In order to evaluate the created model of the machine for collusion detecting, it is urgent to model the practical operating points of the electricity market as test data. To do this, such as a real market, financial-physical interaction between Gencos and system operator is considered in modeling test data process. In other words, in a real market, ISO receives generator's bids (α), then calculates final market data, such as MCP and Q_s, for each generator by solving social welfare optimization problem. However, what is really important in modeling test data is how to model the bidding strategies of the companies. As we know, it is obvious that the bidding strategies of all companies in the oligopoly market are varied, especially in different collusion scenarios. In this section, we assume all generator submit intercept of their supply curve (α) as their hourly optimal bids, then ISO solves the social welfare optimization problem to calculate the final market operation points, such as MCP and Q_s and other collusion criteria, finally uses the trained machine to detect colluding companies. This procedure has been shown in Algorithm 3 and Fig. 1. In order to model and generate test data in collusion scenarios, we have to consider two strategies from Gencoes' viewpoint as follows.

1) To what extent do Gencoes take risks?
2) To what extent are Gencoes risk-averse?

In a preplanned coalition, the conspirators may be risk-taker and adopt high-level bids in their coalition, which satisfy the maximum profits. As we know such strategies have a significant risk to reveal, on the other hand, the conspirators may be

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

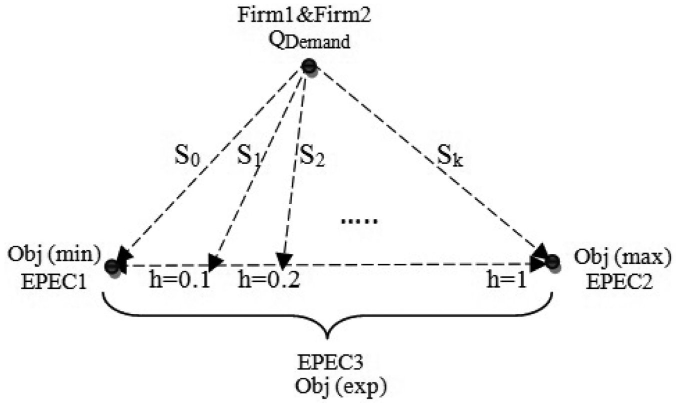

Fig. 2. Strategies of collaborators.

TABLE I
UNITS OWNED BY EACH FIRM

Generation Firms	Units of IEEE test system	
	IEEE 57-bus	IEEE 30-bus
Firm1	U_{11}, U_{12}, U_{13}	U_{11}, U_{12}
Firm2	U_{21}, U_{22}	U_{21}, U_{22}
Firm3	U_{31}, U_{32}	U_{31}, U_{32}

risk-averse in order to hide their coalition and consider lower bids close to the competitive level. In the next section, we consider two aforementioned scenarios and then generated test data will be analyzed by machine learning algorithm separately.

a) Adopting optimal strategy: In this section, we assume conspirators or companies who exercise collusion adopt the highest (optimal) bids in their coalition, which are intercept of their supply curve (α). To do this, it is urgent that united companies estimate the cost function of generator related to rival companies by using Monte Carlo theory. They consider a set of cost function coefficients for each generator of other companies, then consider the average of these sets as the main coefficient of the cost functions. In the other side, ISO after receiving the generator's bids determine market price, power quantities, and calculate collusion criteria. These criteria will be entered into the machine in order to analyze the collusive behavior of all firms. This process is repeated for a different number of load levels and for each possible collusion scenario. In this section, we intended to analyze the efficiency of supervised learning using created test data when conspirators adopt the optimal strategy in their coalition.

b) Adopting nonoptimal strategy: In previous section, it was assumed that conspirators were risk taker and considered optimal amounts (α) as their offers. In a real electricity market, dual behavior of generators and offering the high-level prices can be a sign of market power and collusion, therefore proposed bidding strategy in the previous section can be a risky strategy, accordingly, it is possible that conspirators in order to hide their coalition adopt the special nonoptimal strategy. They come to an agreement on their expected profit first, then are looking for the bid levels that satisfy this agreed profit in their collusion. It can be stated that generators withdraw their highest profits and choose reducing the risk of being detected by ISO instead of maximizing their colluded profit. For example, as Fig. 2, we assume that firms 1 and 2 set out to collude together in special load level QD_i, first they decide to compete separately in an oligopoly environment, then by solving EPEC problem submit their bids to market operator. Summation of profits for two firms is Obj_{min}. For the second time and under similar load demand, companies collude together and solve EPEC jointly. They first decided to adopt the highest profit in their coalition. Therefore consider α by solving EPEC as optimal bids and submit them to market. After gate closure and determining the market price by ISO, they calculate the summation of obtained profits. Therefore, when they collude and select maximum bids or optimal offers, summation profits will be equal to Obj_{max}. For the third time in order to hide their collusion, conspirators are looking for nonoptimal offers/bids and choose their expected profits in the range of $\text{Obj}_{\text{min}} \leq \text{Obj} \leq \text{Obj}_{\text{max}}$ by solving EPEC problem. Therefore, they are looking for bids that satisfy their expected profits. Accordingly in optimization problem (5)–(10), a new constraint is added as follow:

$$\sum_{i \in f,g} \left(\lambda Qs_i - a_i Qs_i - \frac{1}{2} Qs_i^2 \right) = \text{Obj}_{\text{exp}} \qquad (30)$$

where f and g are conspirators and Obj_{exp} is their expected summation profits that is between Obj_{min} and Obj_{max}.

Conspirators can set this value to achieve different level bids from competitive level to maximum bids in collusion as Fig. 2. As a specific strategy, the average of these values can be considered as their expected profits, therefore Obj_{exp} is equal to $\frac{\text{Obj}_{\text{min}} + \text{Obj}_{\text{max}}}{2}$.

It is clear that the more company's expected profit becomes closer to Obj_{min} indicates that conspirators are risk-averse and they behave close to competitive scenario to hide their collusion. Also, the more expected profit becomes closer to Obj_{max}, it means that conspirators are risk taker and consider higher bids in their coalition. Therefore, collusion strategies can be changed by altering collusion coefficient of h linearly, as shown in Fig. 2 and (31)

$$\text{Obj}_{\text{exp}} = \text{Obj}_{\text{min}} + (\text{Obj}_{\text{max}} - \text{Obj}_{\text{min}})h \qquad (31)$$

where $h \in [0,1]$. It is obvious that if $h = 0$ there is no collusion in the market and all Gencoes compete together, also if $h = 1$, conspirators choose maximizing their colluded profit. In the third EPEC problem, we are not looking for objective function optimization, but our main goal is to satisfy the new constraint (30), therefore, the objective function can be a constant value. The result of such optimization problems can be any random value.

V. Simulation and Analysis of Results

In this section, the proposed approach is applied to the IEEE 57 and 30 bus as a test systems. Table I shows the list of generators owned by each producer and Table II shows the list of bilateral collusion scenarios for the studied market.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
DIFFERENT SCENARIOS OF COLLUSION

Scenarios	Collusion type	label
Fair Competition	Collusion-free	1
Firm1&Firm2	bilateral collusion	2
Firm1&Firm3		3
Firm2&Firm3		4

TABLE III
COEFFICIENTS OF a_i FOR ALL GENERATORS FROM ISO'S VIEWPOINT

IEEE test system	Firm1	Firm2	Firm3
30-bus	a_{11}, 17.5	a_{21}, 32.5	30, 30
57-bus	a_{11}, 40, 20	a_{21}, 20	40, 20
30-bus	$17 \leq a_{11} \leq 23$	$7 \leq a_{21} \leq 13$	–
57-bus	$17 \leq a_{11} \leq 23$	$37 \leq a_{21} \leq 43$	–

According to Table II, there are three different scenarios for bilateral collusion. We intend to identify these collusion scenarios from collusion-free scenario and distinguish the conspirators firms. To achieve a proper collusion detection technique, modeling should go beyond the limits of a single period or a certain class of load demand. In this article, we assume load demand varies for 57-bus system between 3000 and 4000 MW and for 30-bus system between 2000 and 3000 MW with 50-MW steps, thus, we have 21 different load levels. Tables VI and VII show parameters of generation units for two IEEE test system.

To create quasi-actual data as training set for machine training from ISO's view, equilibrium points for 21 different load levels and different collusion scenarios are computed according to Table II. As mentioned earlier, first ISO consider a set of cost function coefficient for each generator, then for all different combinations of coefficients, ISO run the EPEC problem in each load level. According to the cost function equation $MC_i = a_i + b_i Qs_i$, we consider b_i as a fixed amount for all sides, also we assume ISO has enough information about the coefficients a_i for all generators as Tables VI and VII except for the first generators of firms 1 and 2 $[U_{11}, U_{21}]$ for two test systems, therefore, as Table III, we assume system operator can estimate upper and lower limits of unknown values.

In Table III, a_{11} and a_{21} are the parameters of generators related to the first unit of firms 1 and 2, respectively, which ISO does not have enough information about its actual amount. Therefore, as Table III from ISO's viewpoint, we select 8 different values in the considered limits for the intercept of marginal cost functions a_i by using normal distribution function. For two IEEE test system by selecting 8 values, we have 64 different scenarios for marginal cost function and for each one, we solve EPEC in each load level then calculate equilibrium points, thus in training process, we generate $64 \times 21 \times 6 \times 4 = 32\,256$ samples for 30-bus system where 4 are the number of collusion scenarios, 6 indicate the number of generators, and 21 are the number of different load levels, also the whole training data for 57-bus system are $64 \times 21 \times 7 \times 4 = 37\,632$ samples. To create final training dataset, collusion criteria is calculated at the generated equilibrium points, thus we have a matrix 32 256 by 5 as a training dataset for 30-bus system and a matrix 37 632 by 5 for 57-bus system, where 5 are the number of criteria. In order to generate the test samples, as mentioned earlier, in collusion scenarios, we consider two approaches from Gencoes' viewpoint: optimal strategies and nonoptimal strategies.

TABLE IV
CONFUSION MATRIX IN THE TEST PROCESS

(a) SVM

		Predicated Class			
		1	2	3	4
Actual Class	1	3220	0	210	245
	2	0	3500	175	0
	3	640	500	2535	0
	4	305	0	80	3290

(b) CART

		Predicated Class			
		1	2	3	4
Actual Class	1	3390	0	130	155
	2	350	3015	270	40
	3	720	555	2400	0
	4	325	45	65	3240

A. Adopt Optimal Strategy

As mentioned in the previous section, all generators and conspirator firms adopt optimal strategies and submit their intercept of supply function (α), which satisfy their maximum profit in their coalition. Therefore, each company in order to solve EPEC problem should estimate the marginal cost of their competitors. To this end, from each company's view, we consider a limit for intercept of marginal cost function (a_i) related to other firms as Tables VIII and IX, then select 10 different values in the considered limits, finally the average of these amounts is considered as actual amount, eventually EPEC problem is solved by firms for 21 load levels and different collusion scenarios. These calculated equilibrium points are submitted as optimal bids/offers by all firms. On the other hand, ISO after receiving the bids, according to Table III for marginal cost scenarios, compute generation powers at equilibrium and then calculate quantities, MCP, and another collusion criteria. It should be noted that in training process, we select 8 different point for marginal cost and we had 64 scenarios, but in evaluating and testing process due to memory limitation of computer system, we consider 5 different values in the considered limits for cost function of two firms, which are unclear for ISO, thus we have 25 different scenarios for marginal costs of generators, therefore, total samples for each collusion scenario are equal to $25 \times 6 \times 21 = 3150$ samples for 30-bus and $25 \times 7 \times 21 = 3675$ samples for 57-bus system

$$\text{Error Rate\%} = \left(1 - \frac{\text{correct predicted samples}}{\text{all samples}}\right) \times 100. \tag{32}$$

Table IV shows the confusion matrix for SVM and CART algorithms in the test process for 57-bus system when conspirators adopt maximum bids in coalition. A confusion matrix contains information about actual and predicted classifications done by the related classification algorithm [35]. The performance or error rates of a learning algorithm is commonly evaluated using the confusion matrix as Table IV and (32) in which correct predicted samples lie on the main diameter. Each row represents the instances in an actual class or target class, whereas each column of the matrix represents the instances in a predicted class

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
PERCENTAGE ERROR OF ALGORITHMS

	SVM		CART		Bag	
	30bus	57bus	30bus	57bus	30bus	57bus
Train	0.2%	6.5%	0.2%	7.44%	0.3%	0.2%
Test	0.9%	14.65%	0.4%	18%	1.3%	14.3%

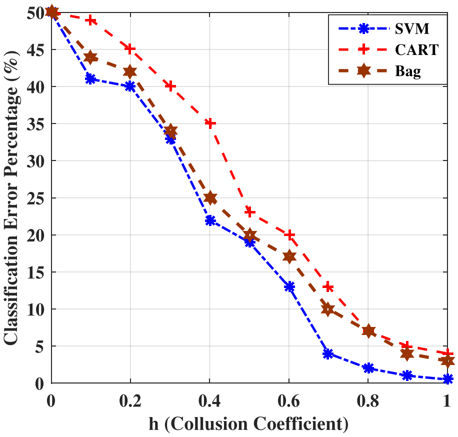

Fig. 3. Classification error of scenarios 1 and 2 (57-bus).

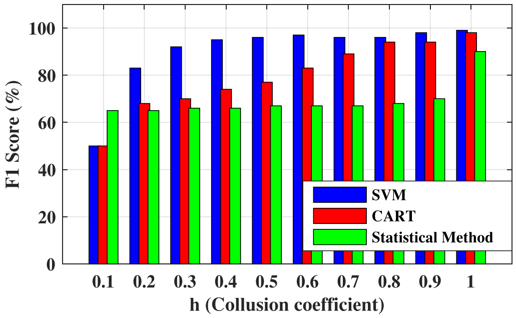

Fig. 4. $F1$ score evaluation for collusion detection.

or output class. For example, in SVM model, of the whole 3675 subsamples of class 2, 3500 samples are correctly predicted to this class and only 175 subsamples have been diagnosed wrongly as class 3. Table V shows the error percentage of algorithms in training and test process for 30- and 75-bus test systems. As it is seen, a four class machine has a proper performance in collusion detection when companies choose maximum bids in their coalitions.

B. Adopt Nonoptimal Strategy

In this section, we analyze nonoptimal strategies of the companies, to do this, we evaluate expected profits from competitive behavior (collusion-free) to maximum profit in collusion. Therefore, we consider a range for profit of conspirators as $\text{Obj}_{\text{min}} \leq \text{Obj} \leq \text{Obj}_{\text{max}}$, then according to (31) select 10 different strategies by choosing 10 different amounts for expected profits by changing h as $h = [0, 0.1, \ldots, 1]$. For each strategy training, test data are generated for different load levels and different collusion scenarios, then efficiency of the machine is analyzed separately. It is clear that if the collusive strategy of companies is closer to competitive behavior/strategy, the machine is confused and cannot properly distinguish between collusion and collusion-free samples, therefore the supervised algorithm has the highest error. Fig. 3 provides a comparison between supervised algorithms in collusion detection and classification of scenarios 1 (collusion-free) and 2 (collusion between firms 1 and 2) for 57-bus system and for all strategies of firms by altering the collusion coefficient h. According to this obtained results, SVM outperforms other supervised algorithms in detection of collusion and has the highest accuracy in classification of two different groups. Furthermore, because for lower values of h, collusion samples are similar to collusion-free samples, machine learning algorithms are confused in classification of these two groups. Fig. 4 compares the performance of supervised algorithms and statistical method in classification of two groups for 57-bus system using F1-score as a precision criterion, which can be calculated as

$$F1 = 2\frac{P_r \times R_e}{P_r + R_e} \tag{33}$$

where P_r and R_e are called precision and recall, respectively, and they are calculated using the following equations:

$$\text{P}_\text{r} = \frac{\text{True positive}}{\text{Predicted Positive}} \tag{34}$$

$$\text{R}_\text{e} = \frac{\text{True positive}}{\text{Actual Positive}} \tag{35}$$

where true positive corresponds to the points that the algorithm detects as positive samples and they are indeed positive ones. Predictive positives are the points that the algorithm detects as positive points but it may have errors. Actual positives are all positive points in the datasets. $F1$ score can never be higher than 1, and the bigger the value of $F1$, the more accurate the classifier in general. As can be seen in Fig. 4, statistical method has a lowest performance even for higher values of h.

Fig. 5 shows the performance of algorithms for classification of scenarios 1 and 2 for the 30-bus test system. According to obtained results, we can see a downward trend in the efficiency of supervised learning algorithms when expected profits of the conspirators become closer to the competitive level. Fig. 6 compares the efficiency of supervised algorithms and statistical method in classification of collusion-free samples from samples of the scenario 2 as Table II for 30-bus test system. According to Fig. 6, SVM has a proper efficiency for collusion detection and classification even for lower values of h. Figs. 7–10 show the classification results of other scenarios for 30-bus and 57-bus test system. According to these figures, SVM model outperforms other supervised methods, also it is clearly seen that the closer the bidding price of companies is to competitive level, the more downward the efficiency of the machine is.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

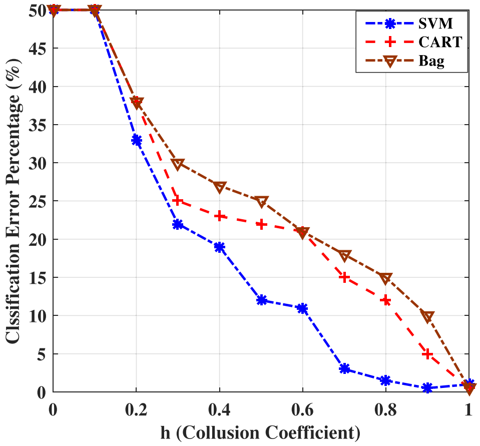

Fig. 5. Classification error of scenarios 1 and 2 (30-bus).

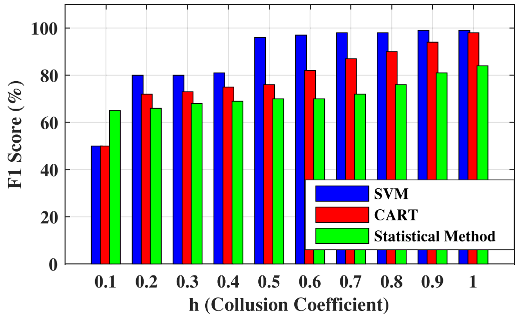

Fig. 6. *F*1 score evaluation for collusion detection.

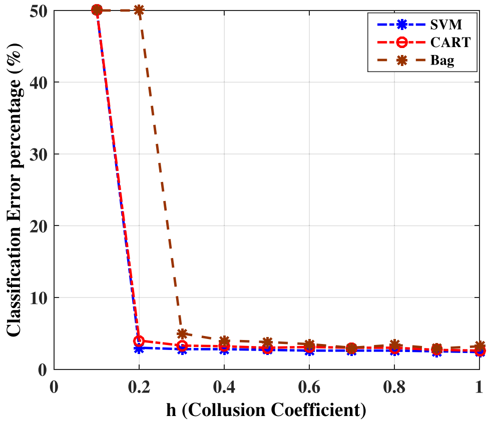

Fig. 7. Classification error of scenarios 1 and 3 (30-bus).

VI. Discussion

Detection of collusion occurrence and identification of colluding companies are the main contributions of this article. One of the barriers to the use of machine learning is the lack of historical and even synthetic data. This article describes how such synthetic data can be created reliably. Due to different load or system changes that occur every day in generation or consumption of electricity, electricity markets do not work on their Nash equilibrium. However, assuming all producers to

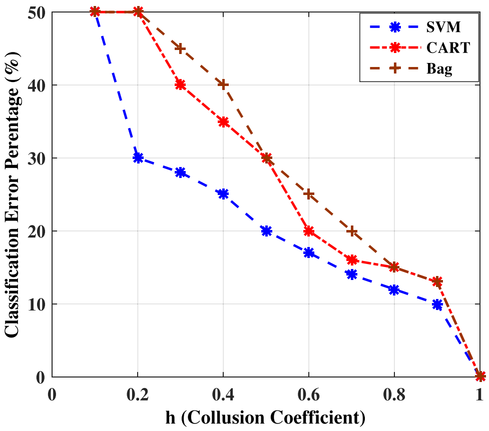

Fig. 8. Classification error of scenarios 2 and 3 (30-bus).

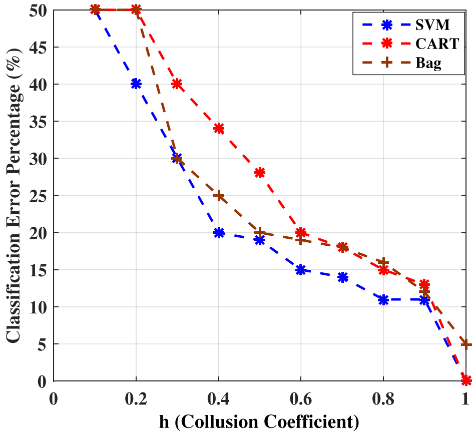

Fig. 9. Classification error of scenarios 1 and 3 (57-bus).

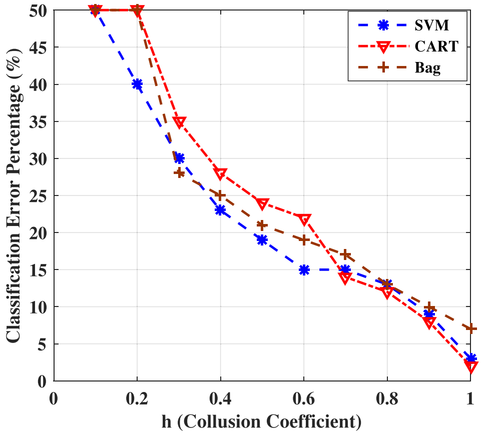

Fig. 10. Classification error of scenario 2 and 3 (57-bus).

be rational market participants, it is reasonable to expect that electricity markets works near their Nash equilibrium. This is why peripheral equilibrium points for creating synthetic data are defined and used as synthetic data for predictive model training in this article. The results and accuracy of collusion detection tools depend on the collusion strategies. Since companies may

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

change their strategies in collusion, as shown in Figs. 3–10, the performance of the machine will be variable, so that the more companies in their alliance adopt higher bids and profits, the better the machine performs.

VII. Conclusion

In this article, an approach for collusion detection in an electricity market is proposed based on Nash equilibrium theory and supervised learning algorithms. Since electricity markets work near their Nash equilibrium, we considered peripheral equilibrium points for creating synthetic data. To this end, first, the possible scenarios of collusion among companies are identified. Then, for each load level and each possible collusion and noncollusion scenarios, market equilibrium is computed by considering uncertainties in marginal cost of all generating units. In collusion scenario, conspirators who are willing to raise the market prices may consider the obtained equilibrium point as their optimal offers, which leads to the highest profit in their collusion, or they may withdraw their optimal offers and change their bids close to the competitive level in order to hide their collusion. Therefore, we examined these two scenarios by applying the proposed approach to the IEEE 57 and 30 buses as test systems and demonstrated that colluding generators choose between maximizing their colluded profit and reducing the risk of being detected by ISO. The results indicated that supervised learning performs very well in detecting collusion and identifying the colluding companies when companies make the maximum profit in their collusion. The accuracy of algorithm, in this case, is over 95%. Moreover, when conspirators withdraw their maximum colluded profit and behave close to the competitive behavior, developed predictive model would struggle to find collusions. Moreover, we showed that the efficiency of SVM algorithm in collusion detection in electricity market is better than other algorithms and statistical methods.

Appendix A
Tables

TABLE VI
Parameters of Generation Units for 30-Bus System

Firms	**unit**	**a**	**b**	Qs^{min}	Qs^{max}
Firm1	1	20	0.2	0	80
	2	17.5	0.175	0	80
Firm2	3	10	0.625	0	50
	4	32.5	0.0834	0	55
Firm3	5	30	0.25	0	30
	6	30	0.25	0	40

TABLE VII
Parameters of Generation Units for 57-Bus System

Firms	**unit**	**a**	**b**	Qs^{min}	Qs^{max}
Firm1	1	20	0.077	0	80
	2	40	0.01	0	80
	3	20	0.25	0	50
Firm2	4	40	0.01	0	55
	5	20	0.022	0	30
Firm3	6	40	0.01	0	40
	7	20	0.032	0	40

TABLE VIII
a_i Values for Generators From Firms Viewpoint for 30-Bus IEEE

From Firm1 viewpoint	From Firm2 viewpoint	From Firm3 viewpoint
$7 \leq a_{21} \leq 13$	$17 \leq a_{11} \leq 23$	$17 \leq a_{11} \leq 23$
$29 \leq a_{22} \leq 35$	$14 \leq a_{12} \leq 20$	$14 \leq a_{12} \leq 20$
$27 \leq a_{31} \leq 33$	$27 \leq a_{31} \leq 33$	$7 \leq a_{21} \leq 13$
$27 \leq a_{32} \leq 33$	$27 \leq a_{32} \leq 33$	$29 \leq a_{22} \leq 35$

TABLE IX
a_i Values for Generators From Firms Viewpoint for 57-Bus IEEE

From Firm1 viewpoint	From Firm2 viewpoint	From Firm3 viewpoint
$37 \leq a_{21} \leq 43$	$17 \leq a_{11} \leq 23$	$17 \leq a_{11} \leq 23$
$17 \leq a_{22} \leq 23$	$37 \leq a_{12} \leq 43$	$37 \leq a_{12} \leq 43$
$37 \leq a_{31} \leq 43$	$17 \leq a_{13} \leq 23$	$17 \leq a_{13} \leq 23$
$17 \leq a_{32} \leq 23$	$37 \leq a_{31} \leq 43$	$37 \leq a_{21} \leq 43$
—	$17 \leq a_{32} \leq 23$	$17 \leq a_{22} \leq 23$

Appendix B
Algorithm

A. Bagging Algorithm

```
Algorithm 2: Bagging.
Input:
    • Training data $S$ with correct labels
      $\omega_i \in \omega = [\omega_1, \ldots, \omega_C]$ representing ; $C$ classes
    • Weak learning algorithm WeakLearn,
    • Integer $T$ specifying number of iterations.
    • Percent (or fraction) $F$ to create bootstrapped
      training data $t = 1, \ldots, T$
  1. Take a bootstrapped replica $S_t$ by randomly drawing
     $F$ percent; of $S$.
  2. Call WeakLearn with $S_t$ and receive the hypothesis
     (classifier) $h_t$.
  3. Add $h_t$ to the ensemble, $E$.
End
Test: Simple Majority Voting- Given unlabeled instance $\boldsymbol{x}$
  1. Evaluate the ensemble $E = \{h_1, \ldots, h_T\}$ on $\boldsymbol{x}$.
  2. Let
       $v_{t,j} = \begin{cases} 1, & \text{if } h_t \text{ picksclass } \omega_j \\ 0, & \text{otherwise} \end{cases}$   (36)
     be the vote given to class $\omega_j$ by classifier $h_t$.
  3. Obtain total vote received by each class
       $V_j = \sum_{t=1}^{T} v_{t,j}, \quad j = 1, \ldots, C$   (37)
  4. Choose the class that receives the highest total vote as
     the final classification.
```

B. Classification And Regression Trees (CART)

```
Algorithm 3: CART Algorithm.
1: Find each feature's best split. For each feature find the
   split, which maximizes the splitting criterion.
2: Find the node's best split among the best splits from
   step i which maximizes the splitting criterion.
3: Split the node using best node split from Step ii and
   repeat from Step i until stopping criterion is satisfied.
```

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

References

[1] S. Stoft, "Power system economics," *J. Energy Literature*, vol. 8, pp. 94–99, 2002.

[2] U. Helman, "Market power monitoring and mitigation in the U.S. wholesale power markets," *Energy J.*, vol. 31, no. 6, pp. 877–904, 2006.

[3] E. S. Amundsen and L. Bergman, "Green certificates and market power on the Nordic power market," *Energy J.*, vol. 33, no. 2, pp. 101–117, 2012.

[4] D. P. Brown and D. E. Olmstead, "Measuring market power and the efficiency of Alberta's restructured electricity market: An energy-only market design," *Can. J. Econ.*, vol. 50, no. 3, pp. 838–870, 2017.

[5] F. M. Mirza and O. Bergland, "Market power in the Norwegian electricity market: Are the transmission bottlenecks truly exogenous?" *Energy J.*, vol. 36, no. 4, pp. 313–330, 2015.

[6] N. Fabra and J. Toro, "Price wars and collusion in the Spanish electricity market," in *Proc. Int. J. Ind. Org.*, vol. 23, no. 3, pp. 155–181, 2005.

[7] R. M. Benjamin, "Tacit collusion in real-time U.S. electricity auctions," USAEE Working Paper 11-085, 2011.

[8] M. Shafie-Khah, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, "Development of a virtual power market model to investigate strategic and collusive behavior of market players,"*Energy Policy*, vol. 61, pp. 717–728, 2013.

[9] M. Shafie-Khah, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, "Exante evaluation and optimal mitigation of market power in electricity markets including renewable energy resources," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 8, pp. 1842–1852, 2016.

[10] A. L. Liu and B. F. Hobbs, "Tacit collusion games in pool-based electricity markets under transmission constraints,"*Math. Program.*, vol. 140, no. 2, pp. 351–379, 2013.

[11] G. J. Werden, "Consumer welfare and competition policy," in *Competition Policy and the Economic Approach*. Cheltenham, U.K.: Elgar, 2011, pp. 11–43.

[12] K. Dawar and P. Holmes, "Trade and competition policy," in*The Ashgate Research Companion to International Trade Policy*. Farnham, U.K.: Ashgate Publishing, 2016, p. 225.

[13] M. Nazrullslam, S. R. Haque, K. M. Alam, and M. Tarikuzzaman, "An approach to improve collusion set detection using MCL algorithm," in *Proc. Int. Conf. Comput. Inf. Technol.*, 2009, pp. 237–242.

[14] G. K. Palshikar and M. M. Apte, "Collusion set detection using graph clustering," *Data Mining Knowl. Discovery*, vol. 16, no. 2, pp. 135–164, 2008.

[15] R. C. Mihailescu and S. Ossowski, "A framework for fraud discovery via illicit agreements in energy markets," *AI Commun.*, vol. 28, no. 4, pp. 607–616, 2015.

[16] D. E. Aliabadi, M. Kaya, and G. Şahin, "Determining collusion opportunities in deregulated electricity markets," *Elect. Power Syst. Res.*, vol. 141, pp. 432–441, 2016.

[17] M. J. Zideh and S. Mohtavipour, "Two-sided tacit collusion: Another step towards the role of demand-side," *Energies*, vol. 10, no. 12, 2017, Art. no. 2045.

[18] B. F. Hobbs, C. B. Metzler, and J. S. Pang, "Strategic gaming analysis for electric power systems: An MPEC approach," *IEEE Trans. Power Syst.*, vol. 15, no. 2, pp. 638–645, May 2000.

[19] M. O. Buygi, H. Zareipour, and W. D. Rosehart, "Impacts of large-scale integration of intermittent resources on electricity markets: A supply function equilibrium approach," *IEEE Syst. J.*, vol. 6, no. 2, pp. 220–232, Jun. 2012.

[20] P. Simon, *Too Big to Ignore: The Business Case for Big Data*, vol. 72. Hoboken, NJ, USA: Wiley, 2013.

[21] R. Kohavi and F. Provost, "Glossary of terms,"*Mach. Learn.*, vol. 30, no. 2/3, pp. 271–274, 1998.

[22] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards, *Artificial Intelligence: A Modern Approach*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[24] V. N. Vapnik, *Statistical Learning Theory*, vol. 1. New York, NY, USA: Wiley, 1998.

[25] K. B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Proc. Int. Workshop Multiple Classifier Syst.*, Jun. 2005, pp. 278–285.

[26] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[27] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data Mining Techniques for the Life Sciences*. Totowa, NJ, USA: Humana Press, 2010, pp. 223–239.

[28] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2003.

[29] M. Esmalifalak, L. Lanchao, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Syst. J.*, vol. 11, no. 3, pp. 1644–1652, Sep. 2017.

[30] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*. Singapore: World Scientific, 2014.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Cole Advanced Books and Software). Monterey, CA, USA: Wadsworth & Brooks, 1984.

[32] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (The Wadsworth and Brooks-Cole Statistics-Probability Series). New York, NY, USA: Taylor & Francis, 1984.

[33] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[34] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, 2009, Art. no. 15.

[35] R. Kohavi and F. Provost, "Glossary of terms,"*Mach. Learn.*, vol. 30, no. 2/3, pp. 271–274, 1998.

Peyman Razmi received the M.Sc. degree in electrical engineering from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2015.

His main research interests include electricity market, machine learning, convex and stochastic optimization, and application of game theory in deregulated power market.

Majid Oloomi Buygi (Member, IEEE) received the Ph.D. degree in electrical engineering from the Darmstadt University of Technology, Darmstadt, Germany, in 2004.

From 2005 to 2008, he was the Director of the Electric Power Engineering Department, Shahrood University of Technology, Shahroud, Iran. In 2009, he was with the University of Calgary as a Postdoctoral Fellow. In 2010, he joined the Ferdowsi University of Mashhad, Mashhad, Iran. He was with the University of Calgary and University of Saskatchewan as a Visiting Professor in 2014 and 2017, respectively. His research interests include power system operation and planning, physical and financial electricity markets, and micro and smart grids.

Mohammad Esmalifalak (Member, IEEE) received the Ph.D. degree from the University of Houston, Houston, TX, USA, in August 2013.

He was a Data Scientist for different industries, such as manufacturing or oil and gas. His specialty is predictive maintenance where he is building predictive models for such industries.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 29,2020 at 12:27:24 UTC from IEEE Xplore. Restrictions apply.