IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 21, 2021, accepted July 17, 2021, date of publication July 20, 2021, date of current version August 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3098628

Real-Time Shill Bidding Fraud Detection Empowered With Fussed Machine Learning

WAJHE UL HUSNIAN ABIDI[1,2], MOHAMMAD SH. DAOUD[3], BAHA IHNAINI[4], MUHAMMAD ADNAN KHAN[5], TAHIR ALYAS[1], AREEJ FATIMA[1], AND MUNIR AHMAD[6], (Member, IEEE)

[1]Department of Computer Science, Lahore Garrison University, Lahore 54792, Pakistan
[2]Digital Commerce Department, Systems Ltd., Lahore 54792, Pakistan
[3]College of Engineering, Al Ain University, Al Ain, United Arab Emirates
[4]Department of Computer Science, Wenzhou-Kean University, Wenzhou 325060, China
[5]Pattern Recognition and Machine Learning Laboratory, Department of Software, Gachon University, Seongnam 13557, South Korea
[6]School of Computer Science, National College of Business Administration and Economics, Lahore 54000, Pakistan

Corresponding author: Muhammad Adnan Khan (adnan@gachon.ac.kr)

ABSTRACT Shill Bidding (SB) occurs when the fake bidders are introduced by the seller's side to increase the final price. SB is a crime committed during the e-Auction, and it is pretty difficult to detect because of its normal bidding behavior. The bidder gets a lot of loss because he pays extra money, and the sellers benefit from shill bidding, so this article proposed a fusion base model. This proposed model is split into two parts training and validation, into 70 and 30 percent. This model has been divided into three sub-modules; the first module, two machine learning algorithms named Support vector machine (SVM), and Artificial neural network (ANN) trained parallel on the same dataset and predicting the bidding fraud. The prediction of these models becomes the input of the fuzzy-based fussed module, and fuzzy decide the actual output based on SVM and ANN predictions. On every bid, it predicts whether the fraud is committed or not. If the bidding behavior is normal, continue the bidding; otherwise, cancel the bid and block the user. The prediction accuracy of the proposed fussed machine learning approach is 99.63%. Simulation results have shown that the proposed fussed machine learning approach gives more attractive results than state-of-the-art published methods.

INDEX TERMS Shill bidding, e-auction fraud, online fraud detection, deep learning model.

I. INTRODUCTION

Virtual Marketplace hosted on the internet is known as the E-auction. It is the process of buying and selling items through online platforms. The bidder bids the item, and the highest bidder is the winner of the item. At the beginning of the auction, bidding starts from the lowest price to a higher price depending upon the buyer's interest.

The history of an auction is found in about 500 B.C when the women and the slaves were sold. In those ages, it was legal by law. In the United States, the auction was started to sell the estates, farms, and slaves, with the growth of technology, the auction was started from the computers, fax, smartphone, and many online platforms, e.g., eBay is the first online auction website started in 1995 in the United States. It is the largest auction site whose net value recorded in 2017 is 1.7 billion US dollars [1].

As in auctions, there is the involvement of money, so it attracts some malicious persons. Shill Bidding (SB) is a prevalent method for auction fraud. In SB, bidding item prices are increased by fake bids. As these are real-time bids, so it's difficult to detect because of their normal resemblance behavior. These moneymakers used different types of SB techniques like as

1) Pre-bidding
2) Post bidding
3) In bidding

SB is the cybercrime, and according to the Internet Crime Report (2001-2009), it is at the top of the list which people report through complaints recorded in IFCC.

Tab. 1 describes the online auction fraud reported from 2001-2009. According to IFCC reports, the 2001 to

The associate editor coordinating the review of this manuscript and approving it for publication was Shenghong Li.

TABLE 1. Internet crime report (2001-2009).

Year	Total Percent of that Year	Reported Fraud in US Dollars
2001	42.8%	~7.6 million
2002	46.1%	~6.6 million
2003	61.0%	~12.7 million
2004	71.2%	~1.62 million
2005	62.7%	~2.58 million
2006	44.9%	~28.0 million
2007	35.7%	~37.9 million
2008	25.5%	~44.5 million
2009	10.3%	~19.9 million

2009 online auction fraud period was at the top of the list. For example, in 2001, total online fraud was reported as $17.8 million while 42.8% was online auction fraud, i.e., approximately $7.6 million. The next year 1 January to 31 December 2002, the total online auction fraud reported is $6.6 million, and further fraud is described in Tab.1.

In the eBay market, in 2001, a total of 40 fake accounts are used by SB for selling art paintings and get paid off $300,000 [1], [2]. In 2007 jewelry seller accused SB of fraud with his employees, and they committed this crime consecutively for four years; in 2007, they earned $400,000 in eBay. In 2010 another person was caught committing the SB fraud, and he paid off the fine of £50,000. Another man used two accounts, the first account is selling an item, and by using the second account, he is fake bidding to increase the cost. This man was fined £5,000 under the newly introduced law [3]. In 2012 "Trade Me" found the SB fraud of vehicles. During the investigation, they found that they had been committing SB fraud for the last year. Trade Me closes this trade by using their platform, and each victim pays the fine of $70,000 [1], [4].

People face too much loss in SB, but they do not leave the eBay site for auction bidding because there is a lower expectation of trust. If trust is the important dimension, then the bidding rate was decreasing with time [5].

In the previous research work, the researchers used the supervised, unsupervised, and semi-supervised simple methods. Some of them use preprocessing techniques to accurate the dataset and go part of them if they successfully optimize the dataset. Some researchers did ANN, SVM, Decision tree naïve Bayes, or some other methods. With time, several SB frauds are accused, so according to the past data and the behavior of SB victims, we are creating a cloud base model that can be integrated into the online platforms and overview the different accounts activity. According to their actions, the proposed model will be able to shortlist the suspicious accounts that can be helpful to overcome the SB.

II. LITERATURE REVIEW

Anowar and Sadaoui [4] detect auction fraud in commercial sites; their model has divided into two parts: offline classification and the other part is online classification. In offline classification, scarp data is collected and then preprocess the auction data. They are using Pattern measurement methods on preprocessing data. After the SB pattern measurement, SB is labeling based on data clustering. Some of the data is imbalanced during this clustering, so to handle the imbalanced data classification optimization tool is used. In the online part of their model, real-time data is collected from the site and preprocessed. After preprocessing, pattern measurement is used and classifies the data. Based on classifying data, the model has decided the bidder's activity is suspicious or not; if yes, then the model verify the fraud detection. In this paper, the authors are using different models in which three models show the best accuracy that is SVM (98.1%), Random Forest (97.1%), and ANN with MLP classifier (97.5%).

Ganguly and Sadaoui [6] devised an online base SVM system for SB fraud detection. To fulfill their purpose, authors apply clustering and labeling techniques and solve the misbalancing learning issues. Once bidding is done, data is collected and applied to the model, and the fraud activity is decided more accurately. The authors create an automatic system because of time handling issues, and the accuracy they achieve is 77.8%.

Alzahrani and Sadaoui [7] proposed the algorithm to optimize the dataset. The author used labeling and clustering techniques to optimize the imbalanced data and use the Hoeffding Tree algorithm on the overSampling and the oversampling algorithm. Their proposed algorithm's overall performance is good, which is 99.7%, 94% under-sampling, and over-sampling, respectively.

Alzahrani and Sadaoui [1] proposed the model in which data is collected through an online eBay site from iPhone 7 device. Collected data is the raw data, so data is preprocessed by using pattern measurement based on matrices. This high-quality data is split date-wise, then data is divided into two parts, i.e., training and validation. The authors used 80% of data in training, and the remaining 20% of data used invalidation.

Anowar *et al.* [8] used hierarchy clustering techniques to split the same type of behavior of data, then applied a semi-automated approach to labeling the normal and suspicious data. In this paper, the authors use three oversampling sampling methods, under-sampling, and hybrid sampling. SVM is used to compare these methods' performance using the 5-K fold and 10-K fold. The best accuracy achieved by this research is 94.0%.

Elshaar and Sadaoui [9] focus the problem on multi-dimensional training data. For this purpose, the authors are using the SSC approach. SSC approach helps in fraud detection with the small amount of data, and skewed class distribution is used with the hierarchical clustering approach to detect anomalies in the dataset. In their statistical testing, the SSC model is separate from the regular and ambiguous bidders, and the overall achieved accuracy is 76%.

Ganguly and Sadaoui [10] focused on the dataset imbalance issue, and after preprocessing the data, the author implemented its dataset into three models that are Naïve Bayes, Neural network, and Decision tree. The author claims that

Naïve Bayes is less sensitive than NN and Decision Tree in data quality. On the other hand, the decision tree is working better than other models on the rebalanced training dataset. The best accuracy achieved by the decision tree is 98%.

Gupta and Mundra [11] proposed a hybrid model, which is a combination of 2 methods one of them is the Prevention method (Authentication Phase) and the other method is the Detection method (Fraud Detection using HMM). The authors divided its model into 2 phases that are the training phase and the detection phase. In the training phase model, create the cluster, identify the bidding habit of bidders, choose the initial probability based on the bidder's habit, and construct a sequence of training data in the last step model. While in the detection phase, auctions are placed, models observe users' behaviors and generate the observation, then calculate the test sequences and decide if the behavior is normal or not. On abnormal behavior, models announce the winner or discard the bid. The problem in this model is that if the model found the abnormal behavior, then there is no method to decide the winner announcement or discard. Thus, there is an ambiguity to take the decision, which may fail the system.

Elshaar and Sadaoui [12] make two new patterns in the dataset. On these patterns, authors create a new high-quality dataset used in a semi-supervised machine learning-based model, which helps to label the multi-dimensional data. Afterward, the authors used oversampling and undersampling methods to use imbalanced class issues. The overall best accuracy achieved is 94% by the classifier named Yasti-J48.

Dong *et al.* [13] proposed an SVM-FDF model for detecting real-time fraud. They implement the spread prominence for a limited marketing scheme to update the credibility when an offer is applied, and fraud sampling is driven using the clustering algorithm. Finally, SVM is applied to each finding and specifies that the transaction is corrupted or fraud. The best accuracy achieved by the SVM-FDF model is 96.8%.

Xiao *et al.* [14] introduces the SSL group method for data handling and an ensemble learning technique to propose a GMDH based GCSSE model. This model involves two stages: first is the training of N base classifiers on the initial training set L with a class label. Then, in the second stage, construct a cost-sensitive GMDH neural network to achieve the selective ensemble classification output for the test set. This model is used on five datasets and gets the best accuracy, i.e., 93.20%.

Elshaar and Sadaoui [15] improve their previous work by fraud classes incorrect predictions. For this purpose, the first attempt the integrate CSL with SSC for fraud detection, then adopt a meta-CSL approach to arrange the cost of miscalculation error, while SSC is trained with imbalanced data. With this CSL+ SSC model, they achieve 99% accuracy.

In previous work, the researcher mainly focused on preprocessing techniques, balancing the data, and using differently supervised and semi-supervised learning models. This article proposed a hybrid supervised learning model that combines three models and some preprocessing techniques to accurately the dataset.

III. DATASET

The dataset is collected from the eBay auction record, in which popular brands' e-auction data is collected from the UCI data repository. The total auction record is 6321, which is used to predict real-time fraud detection.

IV. PROPOSED METHODOLOGY

In this article, the proposed model is a real-time cloud and decision-based fusion model to detect fraud in Shill Bidding (DFM-SB). The proposed DFM-SB model is divided into 2 phases: the training phase and the detection phase on eBay auctions of popular brands. The collected data is in scrap form. First, data Labeling, Imbalance data handling, preprocesses data, and missing average. After preprocessing, this data will be used to train in the ANN and check the trained model achieved the Learning Criteria(LC). Then, on the parallel, train the SVM and check it achieved the LC or not.

When both training algorithms met the LC, then decision level fusion is used with the help of fuzzy logic. If LC is not met, then retrain the model as described in Fig.1. Finally, fuzzy decides the actual output based on SVM and ANN results. This training model is stored in the cloud, and on every e-auction, this model will be used to detect fraud. If a user is found guilty, block the user and discard his auction; otherwise, proceed with the auction.

A. ARTIFICIAL NEURAL NETWORK

In ANN, preprocessed data is divided into two parts: training data and validation data. 70:30 of total data are used in the training and validating Phase. This data is running on the 15 hidden layers of neurons and trains the model.

In ANN, there are 11 input neurons and one output neuron, which have two classes that are normal bidding or SB. Between input and output neurons, 15 hidden layers exist.

The mathematical model of ANN model shell bidding is given below:

Criteria are met, then proceed the model to the next step; otherwise, retrain the model.

In the first layer (input layer), there are 11 input neurons represented as $\breve{\upsilon}1, \breve{\upsilon}2, \breve{\upsilon}3, \ldots, \breve{\upsilon}11$. In the second layer (hidden layer), there are 15 neurons represented as $\zeta 1$, $\zeta 2$, $\zeta 3, \ldots, \zeta 15$ and output is represented as "outϱ" as describe in Fig. 1. The biases are represented as б1 and б2 respectively. To calculate the outð, net$\varrho$and outϱ", which can be calculated from the following Eq's 1, 2, 3, and 4.

$$\text{netð} = \text{б}1 \sum\nolimits_{\gamma=1}^{m} (u_{\gamma \text{ð}} * \breve{\upsilon}) \tag{1}$$

$$\text{outð} = \frac{1}{1+e^{-net\text{ð}}} \quad \text{where ð} = 1, 2, \ldots, \text{n} \tag{2}$$

$$\text{net}\varrho = \text{б}2 \sum\nolimits_{\text{ð}=1}^{n} (\text{ƥ}_{\text{ð}\varrho} * \text{outð}) \tag{3}$$

$$\text{out}\varrho = \frac{1}{1+e^{-net\varrho}} \quad \text{where } \varrho = 1, 2, \ldots, \text{r} \tag{4}$$

The total error "E" can be calculated by using Eq. 5.

$$E = \frac{1}{2} \sum\nolimits_{\varrho} \left(\tau_{\varrho} - out_{\varrho}\right)^{2} \tag{5}$$

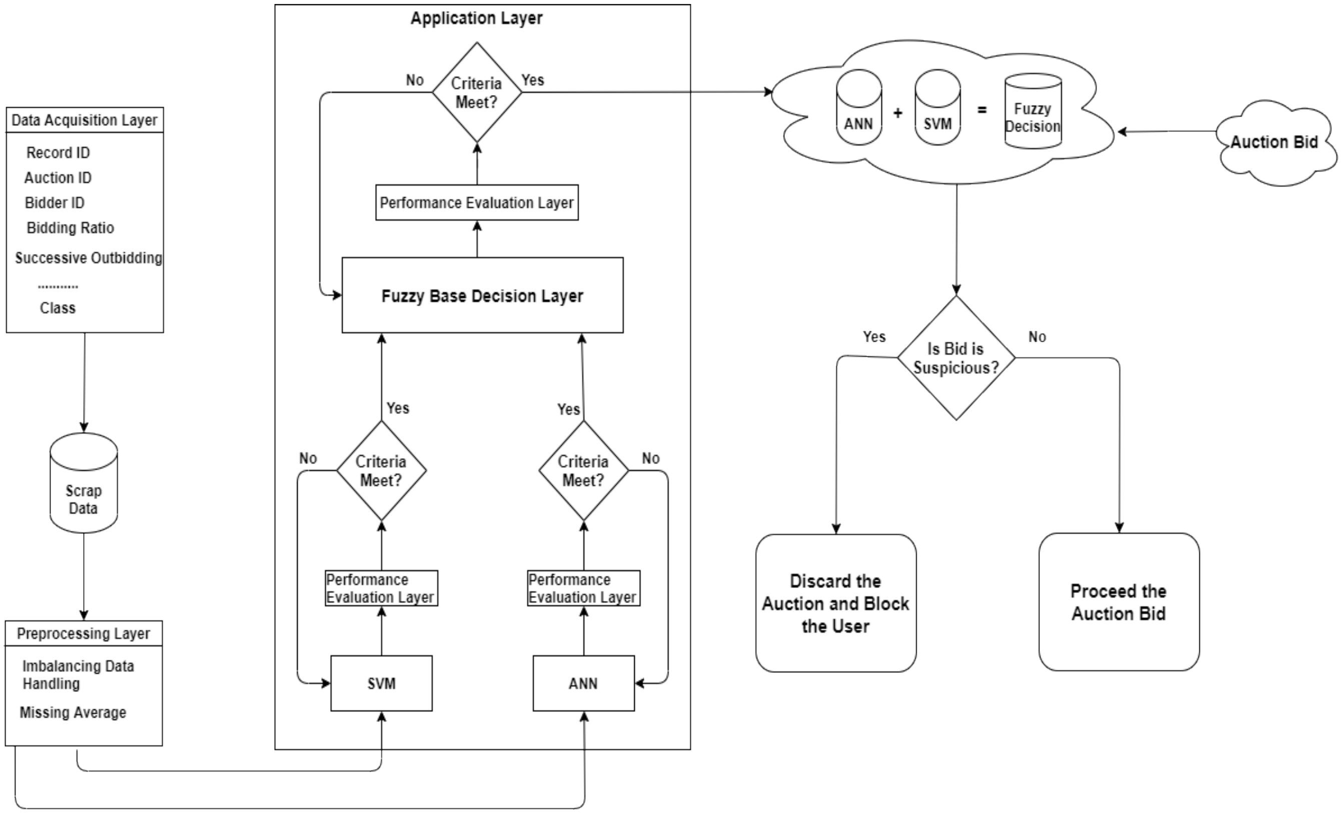


FIGURE 1. Proposed DFM-SB model.

Weights need to be changed concerning errors that can be changed by using Eq. 6

$$\Delta\omega \propto -\frac{\text{ɤ}E}{\text{ɤ}\omega} \tag{6}$$

The weights between the hidden layer and the output layer are updating by using Eq. 7.

$$\Delta \text{ƥ}_{\text{ꝺ},\text{ꝿ}} = -\varepsilon\frac{\text{ɤ}E}{\text{ɤ}V_{\text{ꝺ},\text{ꝿ}}} \tag{7}$$

As $V_{\text{ꝺ},\text{ꝿ}}$ cannot be calculated directly so, calculated it using the Eq. 8 formula.

$$\Delta \text{ƥ}_{\text{ꝺ},\text{ꝿ}} = -\varepsilon\frac{\text{ɤ}E}{\text{ɤ}out_{\text{ꝿ}}} \times \frac{\text{ɤ}out_{\text{ꝿ}}}{\text{ɤ}net_{\text{ꝿ}}} \times \frac{\text{ɤ}net_{\text{ꝿ}}}{\text{ɤ}\text{ƥ}_{\text{ꝺ},\text{ꝿ}}} \tag{8}$$

where $\tau_{\text{ꝿ}}$ is the actual weight of ꝿdescribe in Eq. 9

$$\Delta \text{ƥ}_{\text{ꝺ},\text{ꝿ}} = \varepsilon\left(\tau_{\text{ꝿ}} - out_{\text{ꝿ}}\right) \times out_{\text{ꝿ}}\left(1 - out_{\text{ꝿ}}\right)(out\text{ꝺ}) \tag{9}$$

Eq. 9 is simplified in Eq. 10.

$$\Delta \text{ƥ}_{\text{ꝺ},\text{ꝿ}} = \varepsilon\, \text{ʒ}_{\text{ꝿ}}\, out\text{ꝺ} \tag{10}$$

where value of $\text{ʒ}_{\text{ꝿ}}$ is described in Eq. 11

$$\text{ʒ}_{\text{ꝿ}} = \left(\tau_{\text{ꝿ}} - out_{\text{ꝿ}}\right) \times out_{\text{ꝿ}}\left(1 - out_{\text{ꝿ}}\right) \tag{11}$$

Eq's 12 to 16 are used to update the weights between hidden layer neurons and input layer neurons.

$$\Delta\mu_{\text{ı},\text{ꝺ}} \propto -\left(\sum\nolimits_{\text{ꝿ}} \frac{\text{ɤ}E}{\text{ɤ}out_{\text{ꝿ}}} \times \frac{\text{ɤ}out_{\text{ꝿ}}}{\text{ɤ}net_{\text{ꝿ}}} \times \frac{\text{ɤ}net_{\text{ꝿ}}}{\text{ɤ}out\text{ꝺ}}\right) \times \frac{\text{ɤ}out\text{ꝺ}}{\text{ɤ}net\text{ꝺ}} \times \frac{\text{ɤ}net\text{ꝺ}}{\text{ɤ}\mu_{\text{ı},\text{ꝺ}}} \tag{12}$$

$$\Delta\mu_{\text{ı},\text{ꝺ}} = -\varepsilon\left(\sum\nolimits_{\text{ꝿ}} \frac{\text{ɤ}E}{\text{ɤ}out_{\text{ꝿ}}} \times \frac{\text{ɤ}out_{\text{ꝿ}}}{\text{ɤ}net_{\text{ꝿ}}} \times \frac{\text{ɤ}net_{\text{ꝿ}}}{\text{ɤ}out\text{ꝺ}}\right) \times \frac{\text{ɤ}out\text{ꝺ}}{\text{ɤ}net\text{ꝺ}} \times \frac{\text{ɤ}net\text{ꝺ}}{\text{ɤ}\mu_{\text{ı},\text{ꝺ}}} \tag{13}$$

$$\Delta\mu_{\text{ı},\text{ꝺ}} = \varepsilon\left(\sum\nolimits_{\text{ꝿ}} \left(\tau_{\text{ꝿ}} - out_{\text{ꝿ}}\right) \times out_{\text{ꝿ}}\left(1 - out_{\text{ꝿ}}\right) \times \text{ƥ}_{\text{ı},\text{ꝺ}}\right) \times out_{\text{ꝿ}}\left(1 - out_{\text{ꝿ}}\right) \times \breve{\upsilon}_{\text{ı}} \tag{14}$$

$$\Delta\mu_{\text{ı},\text{ꝺ}} = \varepsilon\left(\sum\nolimits_{\text{ꝿ}} \left(\tau_{\text{ꝿ}} - out_{\text{ꝿ}}\right) \times out_{\text{ꝿ}}\left(1 - out_{\text{ꝿ}}\right) \times \text{ƥ}_{\text{ı},\text{ꝺ}}\right) \times out\text{ꝺ}\,(1 - out\text{ꝺ}) \times \breve{\upsilon}_{\text{ı}} \tag{15}$$

$$\Delta\mu_{\text{ı},\text{ꝺ}} = \varepsilon\left[\sum\nolimits_{\text{ꝿ}} \text{ʒ}(\text{ƥ}_{\text{ꝺ},\text{ꝿ}})\right] \times out\text{ꝺ}\,(1 - out\text{ꝺ}) \times \breve{\upsilon}_{\text{ı}} \tag{16}$$

Eq. 16 can be written in simplified form, as shown in Eq.17.

$$\Delta\mu_{\text{ı},\text{ꝺ}} = \varepsilon\, \text{ʒ}_{\text{ꝺ}}\breve{\upsilon}_{\text{ı}} \tag{17}$$

where value of $\text{ʒ}_{\text{ꝺ}}$, is described in Eq. 18

$$\text{ʒ}_{\text{ꝺ}} = \sum\nolimits_{\text{ꝿ}} \text{ʒ}(\text{ƥ}_{\text{ꝺ},\text{ꝿ}})] \times out\text{ꝺ}\,(1 - out\text{ꝺ}) \tag{18}$$

Weights updating formula describe in Eq. 19.

$$\Delta\mu_{\text{ı},\text{ð}} = \varepsilon\, \text{Ʒ}_{\text{ı}}\, \breve{\text{ʋ}}_{\text{ı}} + \lambda\Delta \text{p}_{\text{ð},\text{g}} \tag{19}$$

Updating weight and hidden layer can be written as in Eq. 20.

$$\Delta\mu_{\text{ı},\text{ð}}(t+1) = \mu_{\text{ı},\text{ð}}(t) + \lambda\Delta\mu_{\text{ı},\text{ð}} \tag{20}$$

After the model is trained, save the training model and validate the model with 30% remaining dataset. Validating data is to enter data into the model and save its results. After saving the results validating data, the output is compared with the actual output, and it achieved 99% prediction accuracy.

B. SUPPORT VECTOR MACHINE

SVM is supervised machine learning and is used in the smaller dataset. The idea behind the SVM is to draw the hyperplane that separates it into different classes. SVM separates the Shill Bidding and normal bidding. To separate the classes in a hyperplane, first, we draw the line. As the equation of a line is described in Eq. 21

$$\chi_2 = a\chi_1 + b \tag{21}$$

where a is the slope of the line and b is the intersect point so that it can be written as

$$a\chi_1 - \chi_2 + b = 0$$

Let suppose $\bar{x} = (\chi_1, \chi_2)^T$ & $\bar{\omega} = (a-1)$ then above equation can be written as Eq. 22

$$\vec{\omega}.\bar{x} + b = 0 \tag{22}$$

This equation is called the equation of hyperplane and is useful for multi-dimensional vectors.

Eq. 23 describe the vector of $\bar{x} = (\chi_1, \chi_2)$ is written as $\bar{\omega}$.

$$\omega = \frac{x_1}{\|x\|} + \frac{x_2}{\|x\|} \tag{23}$$

where $\|x\|$ is defined as

$$\|x\| = \sqrt{x_i^2 + x_2^2 + x_3^2 + \cdots + x_n^2}$$

As we know that the value of $\cos(\varsigma)$ is

$$\cos(\varsigma) = \frac{x_1}{\|x\|}$$

And the value of $\cos(\text{ϐ})$ is

$$\cos(\text{ϐ}) = \frac{x_2}{\|x\|}$$

Now, Eq. 23 can be written the value of ω as

$$\begin{aligned} \omega &= (\cos(\varsigma), \cos(\text{ϐ})) \\ \vec{\omega}\cdot\vec{x} &= \|\omega\|\,\|x\|\cos(\varsigma) \end{aligned} \tag{24}$$

As $\varsigma = \vartheta - \text{ϐ}$, then

$$\begin{aligned} \cos(\varsigma) &= \cos(\vartheta) - \cos(\text{ϐ}) \\ \cos(\varsigma) &= \cos(\vartheta)\cos(\text{ϐ}) - \sin(\vartheta)\sin(\text{ϐ}) \end{aligned}$$

$\cos(\varsigma)$ can also be written as

$$\cos(\varsigma) = \frac{\omega_1}{\|\omega\|}\frac{x_1}{\|x\|} + \frac{\omega_2}{\|\omega\|}\frac{x_2}{\|x\|}$$

By simplifying the above Eq.

$$\cos(\varsigma) = \frac{\omega_1 x_1 + \omega_2 x_2}{\|\omega\|\,\|x\|}$$

Put the value of $\cos(\varsigma)$ is Eq. 24.

$$\vec{\omega}\cdot\vec{x} = \|\omega\|\,\|x\|\frac{\omega_1 x_1 + \omega_2 x_2}{\|\omega\|\,\|x\|}$$

As the above Eq. explain the two dimensions vector, for the n-dimensions vector, it can be written as shown in Eq. 25

$$\vec{\omega}\cdot\vec{x} = \sum\nolimits_{i=1}^{n} \omega_i x_i \quad \text{where i} = 1, 2, \ldots, \text{n} \tag{25}$$

Eq. 25 is used to validate the correctly classifying the data

$$\text{D} = \ddot{\text{y}}(\text{w.x} + \text{b})$$

Given data is correctly classified if the value of D is greater than 0; if not, it is not correctly classified. For our SB data set, compute the dataset onto D for i$^{\text{th}}$ times which can be mathematically represented as

$$D_i = \ddot{y}_i\,(\omega.x + b)$$

d is called the functional margin of the dataset and is written as

$$d = \min_{i=1\ldots m} D_i$$

The hyperplane is selected as favorable, which has the most significant value. Where do is called the geometric margin of the dataset, we find out the optimal hyperplane in this article. To find out the optimal hyperplane, use the Lagrangian function i.e.

$$\begin{aligned} \text{Ɣ}(\omega, b, \text{ϐ}) &= \frac{1}{2}\omega\cdot\omega - \sum_{i=1}^{m}\text{ϐ}_i\,[y : (\omega.x + b) - 1] \\ \nabla_\omega \text{Ɣ}(\omega, b, \text{ϐ}) &= \omega - \sum\nolimits_{i=1}^{m}\text{ϐ}_i y_i x_i = 0 \qquad (26) \\ \nabla_b \text{Ɣ}(\omega, b, \text{ϐ}) &= -\sum\nolimits_{i=1}^{m}\text{ϐ}_i y_i = 0 \qquad (27) \end{aligned}$$

Get from Eq. 26 and 27, we can write as Eq. 28.

$$\omega = \sum\nolimits_{i=1}^{m}\text{ϐ}_i y_i x_i \quad \mathit{and} \quad \sum\nolimits_{i=1}^{m}\text{ϐ}_i y_i = 0 \tag{28}$$

By substituting the Lagrangian function Ɣ

$$\omega(\text{ϐ}, b) = \sum_{i=1}^{m}\text{ϐ}_i - \frac{1}{2}\sum_{i=1}^{m}\cdot\sum_{j=1}^{m}\text{ϐ}_i\text{ϐ}_j y_i y_j x_i x_j$$

Thus, the above Eq. can also be defined in Eq. 29.

$$\max_{\text{ϐ}} \sum\nolimits_{i=1}^{m}\text{ϐ}_i - \frac{1}{2}\sum\nolimits_{i=1}^{m}\cdot\sum\nolimits_{j=1}^{m}\text{ϐ}_i\text{ϐ}_j y_i y_j x_i x_j \tag{29}$$

where i = 1, 2, 3, . . . , m

Due to constraint inequalities, extend the Lagrangian multipliers method to Karush-Kuhn Tucker (KKT) condition, which state that

$$\beta_i \left[y_i \left(\omega_i \cdot x^* + b\right) - 1\right] = 0 \tag{30}$$

Eq. 30 x^* is the optimal point and β is the positive value, and for other points, its values are nearly equal to zero. So, we can write as in Eq. 31

$$y_i \left(\omega_i \cdot x^* + b\right) - 1 = 0 \tag{31}$$

These are the closest points to the hyperplane is also known as support vectors. According to Eq. 31,

$$\omega - \sum_{i=1}^{m} \beta_i y_i x_i = 0$$

This can also be written as

$$\omega = \sum\nolimits_{i=1}^{m} \beta_i y_i x_i \tag{32}$$

Eq. 33 gets when we compute the value of b

$$y_i(\left(\omega_i \cdot x^* + b\right) - 1) = 0 \tag{33}$$

Multiply both sides with y_i

$$y_i^2(\left(\omega_i \cdot x^* + b\right) - 1) = 0$$

As we know y_i^2 is equal to 1

$$b = y_i - \omega_i \cdot x^* \tag{34}$$

$$b = \frac{1}{S} \sum\nolimits_{i=1}^{S} (y_i - \omega.x) \tag{35}$$

In Eq. 35, S is the number of support vectors, and on the hyperplane, we make the predictions.

The hypothesis function is described in Eq. 36

$$U_{SVM} = \mathrm{H}\left(\omega_i\right) = \begin{cases} +1 & if\ \omega.x + b \geq 0 \\ -1 & if\ \omega.x + b < 0 \end{cases} \tag{36}$$

The above points of the hyperplane, i.e., +1 is Shill bidding, and the below point of the hyperplane, i.e., −1, is the no shill bidding or the normal bid of the bidder.

The same dataset was used in the SVM, which was used in the ANN. SVM data is tarin to linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, and Coarse Gaussian SVM with 5-fold cross-validation.

For prediction, define the input parameters and output parameters with the k-fold cross-validation, then run this data on all the SVM models. In this article, run the dataset into 5-fold cross-validation. In 5-fold cross-validation, it divides the data into five numbers or chunks and validates with the next five numbers of chunks. After that, chunks are incremented by 5 to the next. In this method, the data is used for input and as well as output. Finally, SVM separated the classes with the hyperplane and predicted the result.

C. FUSSED ML ALGORITHM EMPOWER WITH FUZZY

Membership functions are used in the fuzzy. Output variables of SVM and ANN are used in the fuzzy as input variables. After defining the membership functions, we define the set of rules based on the membership functions of input and output variables. Based on ANN and SVM, out detection on SB fuzzy will decide whether the bidding is normal or shill bided. Blockage of users, discard the bidding is dependent on the fuzzy decision. The mathematically fuzzy-based decision can be written as

$$U_{ANN} \cap U_{SVM}(ANN, SVM) = \min\left[U_{ANN}\left(ANN\right), U_{SVM}(SVM)\right]$$

where U_{ANN} and U_{SVM} represent the membership function of ANN & SVM, respectively. These statements are relating the core ground for the structure of fuzzy rules.

IF *(ANN is SB) and (SVM is SB)*
Then *(Bidding is SB).*
IF *(ANN is SB) and (SVM is Normal)*
Then *(Bidding is SB).*
IF *(ANN is Normal) and (SVM is SB)*
Then *(Bidding is SB).*
IF *(ANN is Normal) and (SVM is Normal)*
Then *(Shill Bidding is Normal).*

According to the output parameters of ANN and SVM, possible outcome parameters are either normal or SB on both models. So, concerning the fuzzy logic, 4 rule sets are described in the Tab. 2.

The proposed DFM-SB model uses the fuzzy set theory to map input feathers. A fuzzy inference engine is represented as a $\ddot{R}u^e$ which is described as

$$\ddot{R}u^e = \zeta^e \times \varsigma^e \tag{37}$$

$$U_{ANN\cap svm} = U_{ANN(\zeta)} \cap U_{svm(\varsigma)} \tag{38}$$

The rules are then deduced as a fuzzy relation Q_4 as:

$$Q_4 = \bigcup\nolimits_{e=1}^{4} \ddot{R}u^e \tag{39}$$

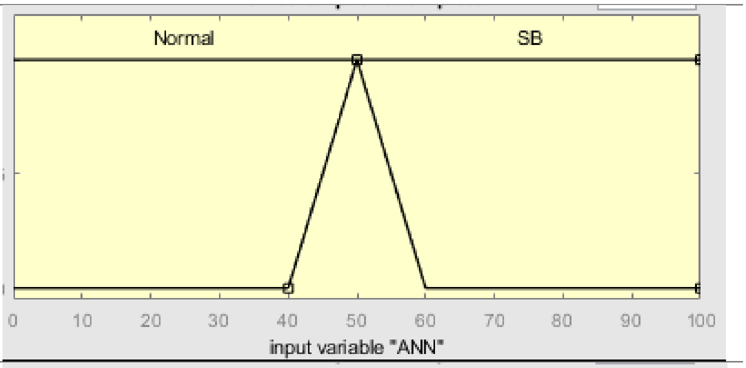
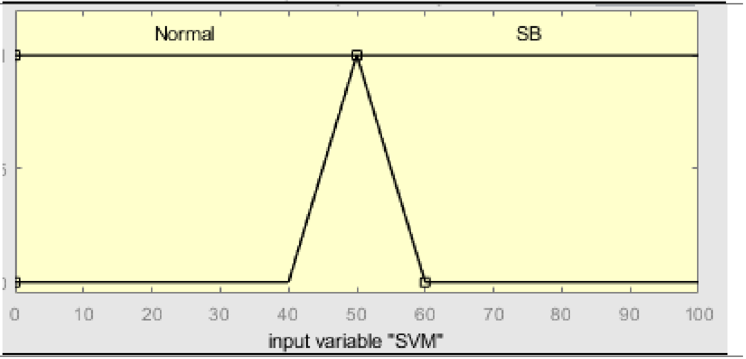
$$U_{\ddot{R}}\left(Decision\ Base\right) = max_{1<x<4}\left[\prod\nolimits_{\gamma=1}^{4}\left(U_{ANN_y}, N_{SVM_y}\right)\right] \tag{40}$$

There are many methods available for defuzzification. De-fuzzifier can be implemented through a centroid method, weighted average, mean-max, and max membership principle. But in proposed model uses the centroid type of defuzzifier. It describes the transformation of fuzzy output generated by the interface engine to frangible using similar functionalities in distinction to those used by the fuzzifier. Eq. (41) describes the crisp point ξ.

$$\xi = \frac{\int \ddot{R}U_{\ddot{R}}\left(\ddot{R}\right) d\ddot{R}}{\int U_{\ddot{R}}\left(\ddot{R}\right) d\ddot{R}} \tag{41}$$

Fig. 2 describes that SVM and ANN are on the x and y-axis of the graph while SB-Detection is on the z-axis. Color-wise, yellow is define the SB detection, while the dark blue area is

TABLE 2. Membership function of proposed DFM-SB system empower with fuzzy.

$ANN_{(U_{ANN(\zeta)})}$	$U_{(Normal)}(\zeta) = \left\{\max\left(\min\left(1,\frac{60-\zeta}{20}\right),0\right)\right\}$ $U_{(SB)}(\zeta) = \left\{\max\left(\min\left(\frac{\zeta-40}{20},1\right),0\right)\right\}$	
$SVM_{(U_{SVM(\varsigma)})}$	$U_{(Normal)}(\varsigma) = \left\{\max\left(\min\left(1,\frac{60-\varsigma}{20}\right),0\right)\right\}$ $U_{(SB)}(\varsigma) = \left\{\max\left(\min\left(\frac{\varsigma-40}{20},1\right),0\right)\right\}$	

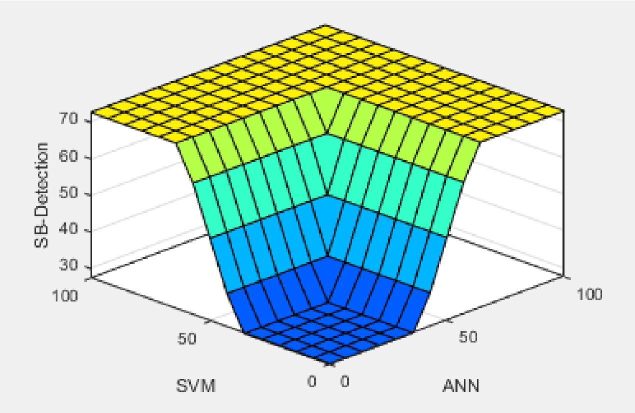


FIGURE 2. Proposed fuzzy decision surface diagram.

TABLE 3. Confusion matrix of ANN (validation).

		Output Results	
Input Values	**Total= 1897**	**Normal Bid**	**SB**
	Normal Bid (1689)	1686	3
	SB (205)	4	204

TABLE 4. Confusion matrix of SVM (validation).

		Outputs Results	
Input Values	**Total= 6321**	**Normal Bid**	**SB**
	Normal Bid (5646)	5605	41
	SB (675)	29	646

defined that the normal bid, while the area between the yellow and blue is may or may not. SB depends upon the rule which we describe in the membership functions.

If the value of SVM and ANN is 0 to 40, then the SB-detection is Normal bid as the value of both models is increased by 40 to 60 then fuzzy decision is in slope between 0 to 1 may be normal bid or SB. When SVM and ANN value is greater than 60, then the SB is detected.

D. RESULTS

In this article, for simulation purposes, MATLAB 2019 has been used. ANN and SVM are used for prediction purposes, and fuzzy is used for decision purposes. The ANN dataset is divided into 70% and 30% ratios of training and testing phase. For validation purposes, there are 1897 number of data is available for which our model predicts the results, which are described in the below Tab. 3.

Tab. 3 describes the actual 1689 bids that are normal in which ANN 1686 truly predicts three wrong predictions. On the other hand, 205 total bids are SB, in which 204 are truly predicted while 4 are wrong predictions.

In SVM, by using 5-k fold cross-validation, there are a total number of 6321 entries, of which 5646 are actual normal bids, and 675 are SB. On SVM, it predicts 5605 truly identified as normal bids while 41 bids are wrongly identified as SB while in actuality, it is normal. 646 are truly identified as SB, and 29 SB are wrongly identified as normal, described in Tab. 4.

TABLE 5. Overall performance of the proposed DFM-SB model.

Overall Performance	Accuracy (%)	Miss Rate (%)
Training	99.8	0.2
Validation	99.6	0.4

TABLE 6. Describe the confusion matrix different parameter values of ANN and SVM and proposed DFM-SB.

Parameters	ANN	SVM	DFM-SB
Accuracy	0.9963	0.9889	0.9963
Miss Rate	0.0037	0.0111	0.0037
Sensitivity	0.9982	0.9927	0.9982
Specificity	0.9808	0.9570	0.9808
PPV	0.9976	0.9949	0.9976
NPV	0.9855	0.9403	0.9855
FPR	0.0192	0.0430	0.0192
FDR	0.0024	0.0051	0.0024
FNR	0.0018	0.0073	0.0018
F1 Score	0.9979	0.9938	0.9979

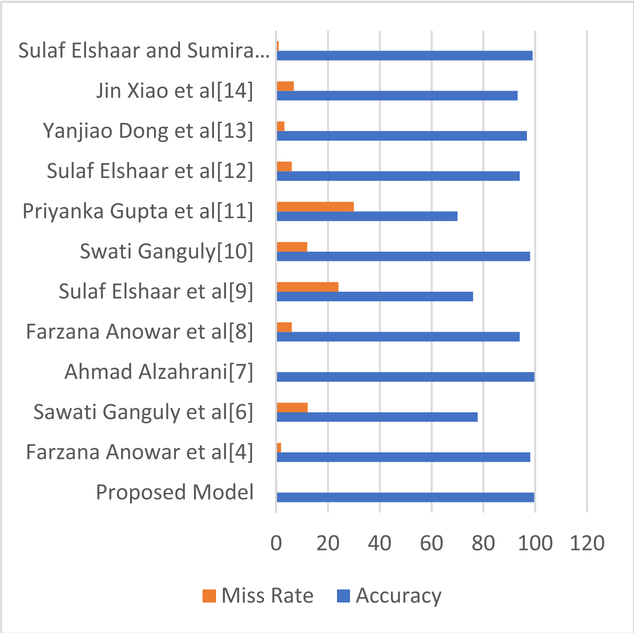


FIGURE 3. Comparison proposed model with the previous model.

Tab.5 lists the overall performance of the DFM-SB model for the training and validation phases. The Proposed DFM-SB model achieved an overall accuracy of 99.8% and a miss rate is 0.2% in training, while the invalidation proposed model achieves 99.6% accuracy and 0.4% miss rate.

Proposed model system performance is measured using the following given fellow statistical formulas. Proposed system accuracy can be measure using Eq. 42

$$\text{Accuracy} = (\text{Tɋ} + \text{Tɲ})/(\text{ɋ} + \text{ɲ}) \tag{42}$$

where Tɋ is the true positive value, Tɲ is the true negative value, ɋ represent the total positive value and ɲ represent the total negative values

$$\text{Miss Rate} = (\text{Fɋ} + \text{Fɲ})/(\text{ɋ} + \text{ɲ}) \tag{43}$$

where Fɋ & Fɲ represents the false positive value and false negative value respectively. Other confusion matrix values are finding by given Eq.'s (44 to 51).

$$Sensitivity = T\text{ɋ}/(T\text{ɋ} + F\text{ɲ}) \tag{44}$$

$$Specificity = T\text{ɲ}/(F\text{ɋ} + T\text{ɲ}) \tag{45}$$

$$PPV = T\text{ɋ}/(T\text{ɋ} + F\text{ɋ}) \tag{46}$$

$$NPV = T\text{ɲ}/(T\text{ɲ} + F\text{ɲ}) \tag{47}$$

$$FPR = F\text{ɋ}/(F\text{ɋ} + T\text{ɲ}) \tag{48}$$

$$FDR = F\text{ɋ}/(F\text{ɋ} + T\text{ɋ}) \tag{49}$$

TABLE 7. Comparison of proposed DFM-SB model with the previous models.

Authors	Best Model	Best Accuracy (%)
Farzana Anowar and Samira Sadaoui[4]	SVM	98.1
Sawati Ganguly and Samira Sadaoui[6]	SVM	77.8
Ahmad Alzahrani[7]	Labeling and Clustering Technique	Under Sampling (99.7) Over Sampling (94.0)
Farzana Anowar, Samaira and Malek Meuhoub[8]	SVM	94.0
Sulaf Elshaar and Sumira Sadaoui[9]	SSC Model	76.0
Swati Ganguly[10]	Decision Tree	98.0
Priyanka Gupta and Ankit Mundra[11]	Hybrid Model	Middle Range Groceries Item (70.0)
Sulaf Elshaar and Sumira Sadaoui[12]	Semi-ML with the help of labelling and Multi-Dimensional	94.0
Yanjiao Dong et al[13]	SVM-FDF	96.8
Jin Xiao et al[14]	GCSSE Model	93.20
Sulaf Elshaar and Sumira Sadaoui[15]	CSL+SSC	99.0
Proposed Model	Fusion base Decision	99.63

$$FNR = F\text{ɲ}/(F\text{ɲ} + T\text{ɖ}) \tag{50}$$

$$F1\ Score = 2T\text{ɖ}/(2T\text{ɖ} + F\text{ɖ} + F\text{ɲ}) \tag{51}$$

Calculate the performance of both ANN and SVM models on our dataset. The results statistic of ANN and SVM are described in Tab. 6 respectively.

The proposed model compares the proposed model in this article and the previous research work described in the literature review section. The proposed model performs a better approach and accuracy than the previous models. Model 3 shows the best accuracy is 99.7, i.e., under-sampling data, and the oversampling results are 94% on average. Our model accuracy is better, and its overall accuracy is 99.6 that is described below Fig. 3.

Tab. 7 described the comparison between the proposed model with the previous models. The authors used different models to achieve the best accuracy. Different data preprocessing techniques are used on SVM and achieve the best 98.1% accuracy by using under-sampling. Oversampling attains 99.7% and 94% accuracy. With the help of the SSC, the model accomplishes 76% accuracy. By using a decision tree successfully achieve 98% accuracy.

In literature, 70% of accuracy was achieved using the hybrid model. However, using labeling and multidimensional preprocessing, the authors attained 94% accuracy while the proposed DFM-SB model accomplishes the accuracy of 99.63%, which is better than the previously published models.

V. CONCLUSION

Online SB detection is a prevalent crime and very difficult to detect because of its very similar behavior during real-time bidding. Due to SB, the other bidder gets a lot of money loss, and the seller receives the extra money. As defined earlier, the previous researcher has a lot of work, but there is always a gap in research. So, this article developed a decision base fusion model used to find the SB in the real-time auction in which SVM and ANN are used for prediction, and Fuzzy is used to decide whether the SB is committed or the bid is normal.

When the bid is made the bidding, the behavior is judge by both SVM and ANN at the same time. Based on their prediction, fuzzy decide the bid is normal or SB.

The proposed model predicts the best results compared to the previous research models, which help detect real-time auction fraud, which helps to block the user and discard fake bids. Therefore, this research will be helpful to both bidders and the e-Auction companies that face the yearly millions of dollars loss and fraud reports.

REFERENCES

[1] A. Alzahrani and S. Sadaoui, "Scraping and preprocessing commercial auction data for fraud classification," Dept. Comput. Sci., Univ. Regina, Regina, SK, Canada, Tech. Rep. CS 2018-05, 2018, p. 17, doi: 10.6084/m9.figshare.6272342.

[2] J. Trevathan, "Getting into the mind of an 'in-auction' fraud perpetrator," *Comput. Sci. Rev.*, vol. 27, pp. 1–15, Feb. 2018, doi: 10.1016/j.cosrev.2017.10.001.

[3] J. Trevathan, C. Aitkenhead, N. Majadi, and W. Read, "Detecting multiple seller collusive shill bidding," Aug. 2018, *arXiv:1812.10868*. [Online]. Available: http://arxiv.org/abs/1812.10868

[4] F. Anowar and S. Sadaoui, "Detection of auction fraud in commercial sites," *J. Theor. Appl. Electron. Commer. Res.*, vol. 15, no. 1, pp. 81–98, 2020, doi: 10.4067/S0718-18762020000100107.

[5] B. C. McCannon and E. Minuci, "Shill bidding and trust," *J. Behav. Exp. Finance*, vol. 26, Jun. 2020, Art. no. 100279, doi: 10.1016/j.jbef.2020.100279.

[6] S. Ganguly and S. Sadaoui, "Online detection of shill bidding fraud based on machine learning techniques," in *Recent Trends and Future Technology in Applied Intelligence* (Lecture Notes in Artificial Intelligence), vol. 10868. Montreal, QC, Canada: Springer, 2018.

[7] A. Alzahrani and S. Sadaoui, "Instance-incremental classification of imbalanced bidding fraud data," in *Proc. 11th Int. Conf. Agents Artif. Intell. (ICAART)*, vol. 2, 2019, pp. 92–102.

[8] F. Anowar, S. Sadaoui, and M. Mouhoub, "Auction fraud classification based on clustering and sampling techniques," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 366–371, doi: 10.1109/ICMLA.2018.00061.

[9] S. Elshaar and S. Sadaoui, "Detecting bidding fraud using a few labeled data," in *Proc. 12th Int. Conf. Agents Artif. Intell. (ICAART)*, vol. 2, 2020, pp. 17–25, doi: 10.5220/0008894100170025.

[10] S. Ganguly and S. Sadaoui, "Classification of imbalanced auction fraud data," in *Proc. Can. Conf. Artif. Intell.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 2017, vol. 10233, no. 2, pp. 84–89, doi: 10.1007/978-3-319-57351-9_11.

[11] P. Gupta and A. Mundra, "Online in-auction fraud detection using online hybrid model," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, May 2015, pp. 901–907, doi: 10.1109/CCAA.2015.7148504.

[12] S. Elshaar and S. Sadaoui, "Semi-supervised classification of fraud data in commercial auctions," *Appl. Artif. Intell.*, vol. 34, no. 1, pp. 47–63, Jan. 2020, doi: 10.1080/08839514.2019.1691341.

[13] Y. Dong, Z. Jiang, M. Alazab, and P. KUMAR, "Real-time fraud detection in e-market using machine learning algorithms," *J. Multiple-Valued Logic Soft Comput.*, vol. 36, nos. 1–3, pp. 191–209, 2021.

[14] J. Xiao, X. Zhou, Y. Zhong, L. Xie, X. Gu, and D. Liu, "Cost-sensitive semi-supervised selective ensemble model for customer credit scoring," *Knowl.-Based Syst.*, vol. 189, Feb. 2020, Art. no. 105118, doi: 10.1016/j.knosys.2019.105118.

[15] S. Elshaar and S. Sadaoui, "Cost-sensitive semi-supervised classification for fraud applications," in *Proc. 12th Int. Conf. (ICAART)*. Valletta, Malta: Springer, 2020, Feb. 2020, pp. 173–187.

WAJHE UL HUSNIAN ABIDI received the B.S. degree in computer science from the University of Sargodha Lahore (UOS). He is currently pursuing the M.Phil. degree with the Department of Computer Science, Lahore Garrison University, Lahore, Pakistan. He is working with Systems Ltd., Pakistan, as a Consultant Ecommerce Services Engineer. He has development experience in different software development domains, such as mobile development (Android and Xambrine), web development (PHP, Asp.Net, react, GraphQL), cloud, image processing, unity 3-D, maya environment designing, and machine learning. His research interests include big data processing, artificial intelligence, cloud computing, and business intelligence.

MOHAMMAD SH. DAOUD received the Ph.D. degree in computer science from De Montfort University, U.K. He is currently an Assistant Professor with the College of Engineering, Al Ain University, United Arab Emirates. His research interests include artificial intelligence, swarm systems, wireless and mobile networks, the Internet of Things, and smart applications.

BAHA IHNAINI received the B.Sc. degree in computer engineering from Philadelphia University and the M.Sc. degree in management information system (MIS) from AABFS, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from Universiti Utara Malaysia, in 2019. He joined the InterNetWorks Research Laboratory during his Ph.D. journey. He worked as a Lecturer with Al Khawarizmi International College, Abu Dhabi branch, from 2011 to 2015. Since August 2020, he has been working as a Lecturer with the Computer Science Department, Wenzhou-Kean University. His current interests include machine learning, sentiment analysis, deep learning, and NLP.

MUHAMMAD ADNAN KHAN received the B.S. and M.Phil. degrees from the International Islamic University, Islamabad, Pakistan, by obtaining a scholarship award from Punjab Information and Technology Board, Government of Punjab, Pakistan, and the Ph.D. degree from ISRA University, Islamabad, by obtaining a scholarship award from the Higher Education Commission, Islamabad, in 2016. Before joining Gachon University, he worked in various academic and industrial roles in Pakistan. He has been teaching graduate and undergraduate students in computer science and engineering for the past 12 years. He is currently working as an Assistant Professor with the Pattern Recognition and Machine Learning Laboratory, Department of Software, Gachon University, Republic of Korea. Presently, he is guiding five Ph.D. scholars and seven M.Phil. scholars. He has published more than 190 research articles with Cumulative JCR-IF and more than 285 in international journals as well as reputed international conferences. His research interests include machine learning, MUD, image processing and medical diagnosis, channel estimation in multi-carrier communication systems using soft computing, and applications of computational intelligence in real-life problems.

TAHIR ALYAS received the M.Phil. and Ph.D. degrees in computer science from the School of Computer Science, NCBA&E, Lahore, Pakistan. Before joining Lahore Garrison University, he worked in various academic and industrial roles in Pakistan. He has been teaching graduate and undergraduate students in computer science and engineering for the past 12 years. He is currently working as an Associate Professor with the Department of Computer Science, Lahore Garrison University, Lahore. Presently, he is guiding ten M.Phil. scholars. His research interests include cloud computing, the IoT, and intelligent age.

AREEJ FATIMA received the M.Phil. and Ph.D. degrees in computer science from the School of Computer Science, NCBA&E, Lahore, Pakistan. She is currently working as an Assistant Professor with the Department of Computer Science, Lahore Garrison University, Lahore. Her research interests include cloud computing, the IoT, intelligent agents, image processing, cognitive machines. She has various publications in international journals and conferences.

MUNIR AHMAD (Member, IEEE) received the Master of Computer Science degree from the Virtual University of Pakistan, in 2018, and the Ph.D. degree from the Computer Science Department, National College of Business Administration and Economics. He has spent several years in industry. He is presently working as the Executive Director and the Head of the United International Group, IT Department, Lahore, Pakistan. He has vast experience in data management and efficient utilization of resources at multinational organizations. His research interests include data mining, big data, and artificial intelligence. He has conducted many research studies on sentiment analysis and utilization of AI for prediction on various healthcare issues.