Received April 11, 2022, accepted May 23, 2022, date of publication May 27, 2022, date of current version June 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178369

An Intelligent Model for Success Prediction of Initial Coin Offerings

MOHAMED GIHAN ALI[1], **ISMAIL IBRAHIM GOMAA**[1], **AND SAAD MOHAMED DARWISH**[2]

[1]Faculty of Commerce, Damanhour University, Damanhour 22514, Egypt
[2]Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, Alexandria 21526, Egypt

Corresponding author: Saad Mohamed Darwish (saad.darwish@alexu.edu.eg)

ABSTRACT Assessment of Initial Coin Offerings (ICOs) is crucial for investment decisions in the ICO market. Since most ICOs succeed in raising funds, failed ICOs must be discriminated against through intelligent classification methods. In this context, this research proposes an intelligent decision model for predicting ICOs' success that merges both the Information Gain Directed Feature Selection (IGDFS) technique as a features rank procedure to select the discriminative features representing the initial pool of features for Genetic Algorithm (GA) to find the ICO's optimal feature set and Fuzzy Support Vector Machine for Class Imbalance Learning (FSVM-CIL) to tackle the problem of imbalanced classification. Two benchmark datasets were used to examine the proposed hybrid model referred to as IGDFS-FSVM. The experimental results reveal that the proposed model that employs an intelligent technique for ICO's feature selection outperforms state-of-the-art classifiers without features selection. In this regard, we conclude that the proposed hybrid model is a practical approach to support investment decisions in the ICO market.

INDEX TERMS Initial coin offerings, fuzzy support vector machine, genetic algorithm, imbalanced classification, feature selection.

I. INTRODUCTION

Initial Coin Offerings (ICOs) or token sales are smart contracts on a blockchain that are used to obtain capital by issuing coins or tokens. A blockchain is an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way [1]. In the case of ICOs, smart contracts are computer protocols on current blockchains that automate transactions of the subsequent form. Essentially, smart contracts replace the intermediary, lowering the transaction costs for firms that need to raise funds [2]. Though similar to crowdfunding in their approach, an ICO's peculiarity is selling tokens. Tokens are cryptographically protected digital units of value that offer benefit to investors in the form of a utility, currency, or security function [3]. For example, utility tokens can be used as a medium of exchange amongst stakeholders on the ICO platform or can be used to buy a product or service in the future. Alternatively, security tokens act as investment resources and qualify their owners for shareholdings, dividends, or alternate financial profits. Additionally, tokens can be traded publicly on a secondary market once the ICO ends.

The significance of ICOs can be attributed to the fact that ICOs have several privileges over traditional funding methods. First, ICOs facilitate obtaining external funds at close-to-zero transaction costs. ICOs enable disintermediation as they are blockchain-based smart contracts. The intermediary's margin can be redistributed as gains for all platform users in the network. Second, the chance to list issued tokens on token exchange platforms shortly after concluding the ICO generates considerable market liquidity. Liquidity helps attract investors that would otherwise avoid investments with long lock-up periods. Moreover, investors can benefit from a token price increase momentarily and can observe the equilibrium price anytime. There is evidence that ICO firms tend to underrate the token so as to attract numerous prospective investors to deepen market liquidity, which will eventually enhance platform success [2]. Third, unlike traditional forms of funding, ICO investors are usually stockholders and customers concurrently. Hence, ICOs assist in capturing future market demand at an initial stage. In addition, they help to generate customer commitment or loyalty if they can gain from token price increases throughout the development stage [4], [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

Despite these advantages, ICOs have turned into a highly controversial issue in the financial world [6]. Due to the absence of regulation, they assist startups to raise funds, while averting compliance costs and intermediaries. Contrarily, being largely unregulated results in an increased investment risk because of malignant behavior, as tokens often do not have current counter-value and do not cause any legal entitlement. For legitimate ventures, they offer equality in crowdfunding, yet the absence of transparency, technical understanding, and legitimacy induces deceitful actors to initiate scam ICO ventures, leading to substantial losses to individual investors and making the market for cryptofinance subject to high-risk.

Although some studies have addressed the determinants of ICO success, they have mostly employed statistical models such as logit, probit, and ordinary least squares regression in their empirical analyses. However, statistical models are subject to several drawbacks. For example, they are likely to be less accurate than machine learning-based models, require a big dataset to attain good performance, are highly sensitive and reactive to multicollinearity, have high sensitivity to outliers, and need dataset to meet specific restrictive assumptions [7]–[9]. In contrast, only a few studies have developed a machine learning- based model to automate the process of ICO success prediction.

Importantly, ICO success prediction involves a considerable number of features. Thus, a solution for dimensionality reduction is necessary before searching for any intuitions in the data. In particular, the objective of feature selection is to address this issue by picking a minor segment of related features from the original large feature set. Through eliminating redundant and immaterial features, feature selection can reduce the dimensionality of the data, accelerate the learning process, make the learned model simpler, and/or boost the performance [10]. Feature selection can be generally categorized into filter and wrapper methods [11]. The filter method is generally computationally efficient and statistically scalable when there is a large set of features under consideration [12]. However, the performance may be low because it doesn't take into account how the classifier interacts with the features and doesn't model how the features depend on each other [13].

Wrapper-based feature selection approaches assess the utility of feature subgroups for a specific classification algorithm.This approach also accounts for the issue of feature dependency. This approach also accounts for the issue of feature dependency. As a result, the wrapper methods typically include a searching procedure for a sound subset of features and thus necessitate computational cost [11], [12]. To decrease the cost of computation, Jadhav *et al.* [14] utilized information gain to direct the feature selection primarily by eliminating features with small information gain, and hence the wrapper method is executed on a reduced space, and the time complexity is declined. Specifically, they applied an information gain directed feature selection algorithm (IGDFS) that ranks features depending on information gain and disseminates the top m features using a genetic algorithm (GA).

Several studies addressing ICO success prediction have utilized imbalanced data (e.g., [15]–[18]). Moreover, Sun *et al.* [9] suggested using imbalanced sampling as the real-world data related to the ICO success prediction topic is imbalanced. However, they suggested using techniques for processing this imbalanced dataset. Handling imbalanced datasets using a standard classifier tends toward the majority data [19], where it assumes that the class distribution of the training data is balanced [20]. Furthermore, the performance measure of the standard classifier is provided with a higher score in the majority class since it is adjusted using the standard accuracy rate. Nevertheless, the significance of the minority class is generally larger than the majority class in imbalanced problems [21]. In addition, minority data is hard to get and, if not properly classified, can result in significant misclassification costs.

Numerous papers have demonstrated the merits of using support vector machine (SVM) in various classification, as it has high prediction accuracy and the capability to handle small datasets while not being constrained by multicollinearity and other statistical assumptions [7]. However, SVM may yield suboptimal results when dealing with imbalanced datasets. Specifically, a standard classification algorithm that considers all training samples uniformly can generate a model that is biassed toward the majority class and has poor results for the minority class. Besides, SVM is vulnerable to outliers and noise in the data.

To address the issue of imbalanced data, the classifiers' performance can be improved through class imbalance learning approaches. These approaches consist of two groups: internal approaches and external approaches. Internal approaches amend the classification algorithms themselves so as to be less sensitive to class imbalance. External methods use the preprocessing of a dataset to make it balanced. Regarding the issue of noise and outliers, Lin and Wang [22] have utilized fuzzy support vector machines (FSVMs), which is a variant of the SVM, to deal with the issue of outliers and noise. However, FSVM can still be sensitive to the problem of class imbalance like SVM. Significantly, Batuwita and Palade [23] suggested a technique to upgrade FSVMs for class imbalance learning (FSVM-CIL), that can be utilized to deal with both the problems of class imbalance and outliers/noise. In their suggested method, they assign fuzzy-membership values to training examples so as to decrease the influence of both of the previously mentioned problems under the cost-sensitive learning principle.

A. MOTIVATION

The first motivation of this paper is to identify ICO success determinants that can decrease the large investment risk borne by investors using a careful evaluation of numerous characteristics, such as the venture's source code, white paper, and social media. Identifying these determinants and their influence on the ICO's success enables investors to evaluate ICOs more precisely despite the considerable uncertainty that surrounds them, leading to better- informed decision- making.

The second motivation for this paper is to develop a hybrid machine learning-based model to bypass the issues associated with statistical models. A machine learning–based model can add value to the cryptocurrency community in two facets. First, it would bridge the gap in prior literature regarding the design of reliable, automatic, and difficult-to- manipulate systems to analyze and assess the performance level of ICO projects. Second, a well-designed machine learning model can provide early warning signs by integrating diverse kinds of information regarding ICOs. Therefore, it can potentially assist investors to identify fraudulent ICO ventures and make rational investments in ICOs.

B. CONTRIBITION

Accordingly, there are numerous contributions to this research. Firstly, the theoretical contribution of this research is to determine ICO success factors by reviewing related literature. Secondly, the practical contribution is through offering a mechanism to automate ICO success prediction, utilizing a hybrid machine learning model that is anticipated to outperform statistical models. Finally, the technical contribution of this paper is to further develop the prediction model of ICO success via combining IGDFS with FSVM-CIL. This integration is predicted to promote better informed decision making.

The remainder of this research is arranged as follows. Section II provides a brief overview of key related works; Section III describes the proposed hybrid machine learning model; Section IV discusses and analyses the empirical findings; and Section V concludes the research.

II. RELATED WORK

Through low transaction charges similar to crowdfunding [24], ICOs have become a substantial driver for financial inclusion by unrestricted access to investments and funds. Moreover, blockchain technologies can lower the entrance barriers for new actors. In spite of the growing number of ICOs in the last few years, there is an absence of a good understanding of the determinants of their success. Such an understanding is critical for ICOs to plan their blockchain-based funding initiatives appropriately and enables prospective investors to search for major signs and driving factors of current ventures. Furthermore, it can aid market participants and regulators to understand how the present regulatory framework is implemented on ICOs. To fill these knowledge gaps, this paper reviews related literature on the success factors of ICOs. Specifically, Table 1 demonstrates the features determining ICO success based on related ICO studies.

However, financial data commonly encompasses features that are redundant and irrelevant [25]. The redundancy and the deficiency in the dataset can decrease the predictive accuracy and result in incorrect decisions [12], [26]. Feature selection eliminates irrelevant and redundant features from data, hence increasing the classification accuracy and decreasing the cost of computation [27], [28]. Due mainly to the huge search space, feature selection is becoming a more challenging task. Thus, an exhaustive search for the optimal feature subset of a particular dataset is practically unfeasible in the majority of situations. Various search methods have been implemented in feature selection; for instance, complete search, greedy search, heuristic search, and random search. Nonetheless, many existing feature selection techniques are still subject to stagnation in local optima and/or great computation cost [10]. Therefore, an efficient global search method is required to better deal with feature selection issues. Lately, great attention has been given by the feature selection community to evolutionary computation approaches as they are recognized for their global search capacity.

More specifically, wrapper-based GA has turned out to be the most established approach employed in the field of feature selection, and it has revealed its efficiency in numerous fields thanks to its capability to find good solutions for complex searching and optimization problems [10], [14]. Nevertheless, the interpretation capacity of the prediction power of every feature in the data is usually a requirement in some applications, such as ICO investments. In these circumstances, a feature selection technique such as information gain that returns a score, is superior to techniques that return only a rank or a feature subset, where the feature significance is not considered [29].

Toward that end, the authors in [14] presented a new technique for feature selection in credit assessment applications that carries out the feature rank based on information gain and propagates the top m features for the GA procedure to select the optimal set. Thier approach uses information-based ranking of features to decrease the feature set by modifying the preliminary population pool of GA so that the best individuals are picked. Furthermore, this measure is used to guide the evolution of GA by modifying the GA parameters of population pool, crossover, and mutation.

Notably, only a few models have been developed for the successful prediction of ICOs. For instance, a number of studies [16], [18], [30], [31] analyzed the features explaining the funding success of token offerings using logit models; while other studies [38], [39], [42] applied probit regressions to examine features predicting the success of ICOs; and Perez, Sokolova, and Konate [43] attempted to explain the role of digital social capital in ICO success by applying an exploratory factor analysis to leverage the key latent features and using structural equation modeling to test the hypotheses and evaluate the model performance. Cerchiello, *et al.* [35] used random forest to emphasize the more relevant predictors and employed linear regression (LR) to differentiate between successful, failed, and fraudulent ICOs.

As the ICO's success prediction is similar to corporate distress prediction and credit scoring, consequently, literatures on the classification algorithms' progress in these

TABLE 1. Description of ICOs features.

Category	Features	Type	Description	Ref.	Data set
ICO Characteristics	Token services	Indicator	Whether the token can be utilized to pay for or access products or services.	[17][30][31]	1
	Token profit	Indicator	Whether the token grants profit rights for their holders	[30][31]	1
	Is cryptographic token	Indicator	Whether the ICO takes the structure of a smart contract on a standing blockchain (e.g., Ethereum)	[30][32-34]	1
	Is currency or general-purpose blockchain	Indicator	Whether the token is planned to be utilized primarily as a currency, replacing traditional fiat money, or as the unit of account for a new general-purpose blockchain capable of executing smart contracts.	[30]	1
	Github	Indicator	Whether the ICO is represented on Github	[16][30] [31][35-37]	1,2
	Smart contract code available	Indicator	For tokens taking the form of a smart contract on another blockchain, whether the smart contract source code is available on Github prior to the ICO	[30]	1
	Pre-ICO/pre-sale	Indicator	Whether the ICO has a pre-ICO	[30][31] [36] [38]	1
	Length of ICO (calendar days, actual)	Discrete	The actual length of the period of crowd sale in days	[30] [34] [39] [40]	1
	Length of ICO (calendar days, planned)	Discrete	The planned maximum length of the period of crowd sale in days	[30]	1
	Pre-sale tokens locked up	Indicator	Whether a lock-up period exists for tokens sold at the stage of pre-sale	[30] [33]	1
	Team tokens locked up	Indicator	Whether some portion of the tokens held by the issuing company and/or the founding team are subject to a vesting program	[30]	1
	Team lockup Period (weighted avg.)	Continuous	The weighted mean of maturity of tokens controlled by the issuing firm as well as the founding team. It also includes all tokens contained in "Token Share Team".	[16] [30]	1
	Know-Your-Customer (KYC)	Indicator	Whether the project used a KYC process where participants had to show that they were who they said they were by giving personal documents.	[30][36][38]	1
	Qualified investors only	Indicator	Whether merely investors with an accredited investor position or equivalent are permitted to take part in the ICO	[30]	1
	Legal form and jurisdiction	Indicator	Whether the venture sponsors have defined the form of legal entity and the relevant jurisdiction for the ICO, thereby providing prospective funders with the bare minimum of legal protection in the event of a scam.	[30] [31]	1
	Legal form is foundation	Indicator	Whether the issuing firm is a not-for-profit foundation	[30]	1
	Legal entity is corporation	Indicator	Whether the issuing firm is a joint-stock corporation or its equivalent in non-US jurisdictions	[30]	1
	Legal entity is limited liability corporation (LLC)	Indicator	Whether the issuing firm is an LLC, limited liability partnership, or their equivalent in non-US jurisdictions.	[30]	1
	Investors have governance rights	Indicator	Whether token holders possess a voting right on investment, business, or governance decisions, including advisory votes	[30]	1
	Postal address known	Indicator	Whether the physical postal address of the ICO promoter's headquarters is known	[30]	1
	Years since foundation	Discrete	Years from the time when the founding team started working on the venture for which the ICO is being conducted. Where unavailable, the incorporation date from the commercial register is used.	[30]	1
	Issuer has customers for product	Indicator	Whether the product or service underlying the ICO has consumers (regardless of whether they pay for the service or not)	[30]	1
	Product can be tried out	Indicator	Whether prospective contributors are able to try the product or prototype	[30]	1
	Product or prototype developed	Indicator	Whether there has been a development of the product for which capital is being raised or an initial "alpha" or "beta" version of it	[30]	1
Disclosure Characteristics	Development road map available	Indicator	Whether the documentation contains a road map with dates and milestones for the development and commercialization of the product	[30][40]	1
	Funding milestones	Indicator	Whether the conditions of funding lay out binding milestones (e.g., improvement of a working prototype) that must be achieved for the capital raised in the ICO to be released to the firm	[30, 40]	1
	Business model available	Indicator	Whether the documentation details the market opportunity the product offers financed through the ICO addresses and how the firm will ultimately earn money.	[30]	1
	Celebrity endorsement	Indicator	Whether the ICO is being promoted via a famous entertainment or sports personality on social media.	[30]	1
	Legal advisor disclosed	Indicator	Whether the legal expert (whether a firm or an individual) who advised the firm in arranging its ICO is disclosed.	[30]	1
	Whitepaper	Indicator	Whether the ICO venture has a whitepaper	[15][17][35]	2
	Whitepaper Length	Discrete	Number of pages in the white paper document	[16][18] [36]	1
Social Media	Telegram	Indicator	Whether the ICO is represented on Telegram	[15][35][36]	2
	Twitter	Indicator	Whether the ICO has a twitter account	[15][35]	2

TABLE 1. ***(Continued.)*** **Description of ICOs features.**

	Facebook	Indicator	Whether the ICO has a Facebook account	[35]	2
	YouTube	Indicator	Whether the ICO is represented on YouTube	[35]	2
	Slack	Indicator	Whether the ICO is represented on Slack	[35]	2
	Reddit	Indicator	Whether the ICO is represented on Reddit	[35]	2
	Bitcointalk	Indicator	Whether the ICO is represented on Bitcointalk	[35]	2
	Medium	Indicator	Whether the ICO is represented on Medium	[35]	2
	# Telegram	Discrete	Number of users in Telegram	[35]	2
	Telegram chat Sentiment Score _ BING	Discrete	Sentiment score depending on the number of match between a predetermined list of positive and negative terms and words included in every text source calculated using BING dictionary.	[35]	2
	Positive words_ BING	Discrete	The total number of positive words in BING dictionary	[35]	2
	Negative words_ BING	Discrete	The total number of negative words in BING dictionary	[35]	2
	Telegram chat Sentiment Score_ NRC	Discrete	Sentiment score depending on NRC dictionary Word-Emotion Association Lexicon	[35]	2
	Positive words_ NRC	Discrete	The total number of positive words in NRC dictionary	[35]	2
	Negative words_ NRC	Discrete	The total number of negative words in NRC dictionary	[35]	2
Financial Details	Tokens supply	Continuous	The total number of tokens	[30] [32-34] [36] [38]	1
	Token supply is fixed	Indicator	Whether the total number of tokens stays fixed indefinitely, as opposed to tokens that permit inflation or the generation of extra tokens under specific conditions	[30]	1
	Token share crowd sale investors	Continuous	If the crowd sale sells out, a portion of the total token supply will be assigned to crowd sale investors after the crowd sale.	[30][36] [39]	1
	Token share pre-sale investors	Continuous	If the crowdsale sells out, a portion of the total token supply will be allocated to pre-sale investors.	[30] [36] [39]	1
	Token share producers/ miners (Incentive pool)	Continuous	A portion of the total token supply is set aside to compensate "miners" or producers on the platform subsequent to the crowd sale, supposing the crowd sale sells out.	[17] [30] [33]	1
	Token share team	Continuous	Portion of total token supply under control of the issuing firm and the founding team subsequent to the crowdsale, supposing the crowdsale sells out. Includes all tokens controlled by the entity, including tokens reserved promotional activities, “bounties” (compensation for promotional activities), compensation of suppliers, employees and advisors and any further residual categories.	[30] [33]	1
	USD price of token	Continuous	How much is one token in the crowdsale worth in US dollars (USD)?	[30][33] [35][40]	1,2
	Crowdsale max. discount	Continuous	Maximum discount given to (usually large or early) investors throughout the stage of crowdsale.	[30]	1
	VC-Backed	Indicator	Whether the ICO is supported via a Venture Capital (VC) fund prior to or during the ICO	[17][30] [38][41]	1
	Crowdsale is auction	Indicator	Whether the price of a token for crowdsale investors relies on the total amount of capital raised throughout the crowdsale	[30]	1
	Use of proceeds disclosed in detail	Indicator	Whether the issuer provides a comprehensive analysis for the use of capital raised during the ICO (e.g., X software developers at Y dollars per hour are required to complete the product in Z hours).	[17][30] [42]	1
	Use of proceeds mentioned	Indicator	Whether the issuer provides a rough breakdown for the use of funds raised during the ICO (e.g., 40% product development, 10% legal, 50% marketing).	[17][30][42]	1
	ICO Hard Cap	Indicator	Whether the ICO had a hard cap	[30] [36]	1
	Unsold tokens “burnt” or proportional allocation	Indicator	Whether unsold tokens are either destroyed or the token allocation is done proportionally (e.g., the team receives 20% of all tokens created following the crowdsale, regardless of its result).	[30]	1
	Unsold tokens kept by issuer	Indicator	Whether the issuer retains unsold tokens, either for sale in the future or to be utilized for a different purpose.	[30]	1
Team Characteristic	Team size	Discrete	Total number of full-time team members in the course of the ICO, excluding advisors and contractors.	[16][30] [35][36] [39][40]	1,2
	Advisors	Indicator	Whether the team includes advisors	[35]	2
	# Advisors	Discrete	Total number of advisors	[15][35] [36]	2
	Experienced team	Indicator	Whether the founding team possess an experience not less than ten years, on average, in management, technology or entrepreneurship	[30]	1
	Team experience missing	Indicator	Whether the information is insufficient to specify the value of the feature “experienced team”	[30]	1
	High-quality advisory team	Indicator	Whether the advisory team is of high quality, i.e., mostly comprised of individuals with substantial expertise as entrepreneurs, executives, venture investors or academics	[30]	1
	Financial advisor disclosed	Indicator	Whether the financial/blockchain expert (either a firm or an individual) who advised the firm in arranging its ICO is disclosed.	[30]	1
	Team member with business background	Indicator	Whether, at the minimum, one of the team members has a substantial expertise in entrepreneurship, consulting or management	[17] [30]	1

TABLE 1. ***(Continued.)*** **Description of ICOs features.**

	Team business background missing	Indicator	Whether there is insufficient information to specify the value of the feature "team member with business background"	[30]	1
Cryptocurrency Dynamics	Investors from other jurisdictions excluded	Indicator	Whether investors from jurisdictions excluding the US are not permitted to take part in the ICO	[30]	1
	Registered in offshore financial center	Indicator	Whether the jurisdiction, in which the company is incorporated, is an offshore financial center as defined by the international monetary fund	[30]	1
	Simple agreement for future tokens	Indicator	Whether the ICO uses a "Simple Agreement for Future Tokens" (SAFT) under which tokens are merely issued as soon as the platform on which they can be utilized has been released	[30]	1
	US retail investors excluded	Indicator	Whether the non-accredited investors from the United States are not permitted to take part in the ICO	[30]	1

fields is a promised model for the development of ICO an success prediction model. A substantial concern that relates to these areas is the implementation of imbalanced classification. As an example, Sun *et al.* [44] used the synthetic minority over-sampling technique (SMOTE) and the bagging ensemble to predict financial distress based on imbalanced data. They used the SVM as the base classifier for this.

In this regard, SVM is an extensively used machine learning algorithm that has been fruitfully employed for numerous real-world classification issues in several domains, because of its strong mathematical background, high generalization power, and capacity to locate global classification solutions [45], [46]. Nevertheless, it has been well-studied that SVM can be susceptible to class imbalance [47]–[50], i.e., the separating hyperplane of the SVM algorithm, developed using imbalanced data, can be skewed toward the minority (positive) class [48], [49]. This skewness generally results in the generation of a large number of false-negative predictions, which lowers the model's predictive power on the positive class in comparison to the predictive power on the negative (majority) class. As mentioned earlier, there are class imbalance learning (CIL) approaches that can be implemented when developing SVM classifiers with imbalanced datasets in order to decrease the influence of data imbalance. Largely, these techniques can be divided into two categories: external techniques and internal techniques [23], [50].

External techniques are independent from the learning algorithm being applied, and they include preprocessing of the training data to balance them prior to training the classification algorithms. Various resampling techniques, such as random and focused oversampling and undersampling, are located in this group. For instance, SMOTE is an oversampling method where new synthetic instances are created in the neighborhood of the current minority-class instances instead of directly replicating them [51]. Internal methods handle modifications of a learning algorithm to make it less susceptible to the class imbalance. For example, Veropoulos *et al.* [47] have suggested an approach called different error costs (DEC), where the SVM objective function has been modified to allocate two misclassification costs; through allocating a greater misclassification cost for the positive (minority) class instances than the negative (majority) class instances in order to reduce the influence of class imbalance.

Recentlty, some significant research has been conducted toward the integration of fuzzy logic and SVMs in diverse ways, and the expression fuzzy SVMs (FSVM) has been used to point to the majority of those different approaches. Lin and Wang [22] applied a fuzzy membership value for every training instance that is based upon the significance of that instance in its class and reformulate the SVM classifier so that different input points can make different contributions when locating the separating hyperplane. Wang *et al.* [52] added to this method by giving two membership values for each training example. These values determine which positive and negative classes the example belongs to.

Moreover, there is a considerable amount of research on how to allocate the fuzzy membership of FSVM. Dai [53] evolved a fuzzy penalty in accordance with the distance function and decaying function, and integrated a total margin algorithm with SVM based on the penalty. Lee *et al.* [54] altered weights in AdaBoost with a weak learner of weighted SVM. The nearest neighbor concept has also been applied to deal with the class imbalance problem. Ando [55] employed class-wise weighting with nearest neighbor density estimation and learned its weight parameters by convex optimization. Fan *et al.* [56] suggested an entropy fuzzy support vector machine (EFSVM) to deal with the issue of class imbalance with the nearest neighbor entropy. Cho *et al.* [57] introduced an instance-based entropy fuzzy support vector machine (IEFSVM) that merges nearest neighbor entropies that change in accordance with neighborhood size for every data point, such that it can allocate the fuzzy membership by reflecting all information of every instance efficiently.

Although the current CIL techniques present for SVMs can overcome the class imbalance issue, they can still be prone to the problem of outliers and noise. On the other hand, FSVM can deal with the problem of outliers/noise; but it can also suffer from the class imbalance problem. Batuwita and Palade [23] presented an improvement to FSVM, called FSVM-CIL, to surpass the issue of class imbalance. In FSVMs, different membership values (or weights) can be allocated to training instances in order to reflect their different importance. Besides, they revealed that this is similar to assigning different misclassification

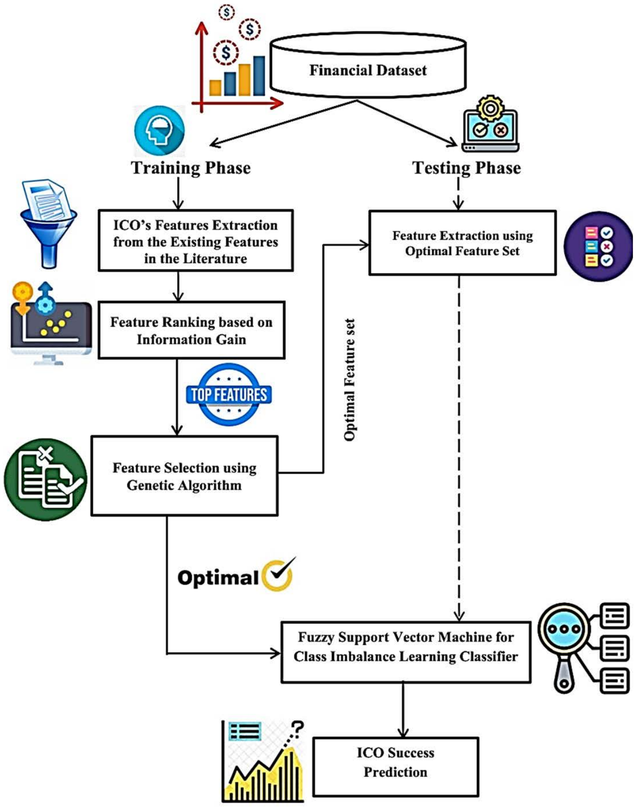
Financial Dataset
Training Phase
Testing Phase
ICO's Features Extraction from the Existing Features in the Literature
Feature Ranking based on Information Gain
TOP FEATURES
Feature Selection using Genetic Algorithm
Optimal Feature set
Feature Extraction using Optimal Feature Set
Optimal
Fuzzy Support Vector Machine for Class Imbalance Learning Classifier
ICO Success Prediction

FIGURE 1. Block diagram of the proposed system.

costs for different training examples. Therefore, to reduce the impact of class imbalance, higher misclassification costs or higher membership values are allocated to the positive-class examples, whereas lower misclassification costs or lower membership values are allocated to the negative-class examples.

III. METHODLOY

In this section, we first outline the dataset employed in this study. Then, we introduce a hybrid model for investment decisions in the ICO market. The block diagram that briefs the key constituents of the suggested hybrid model is portrayed in Fig.1. The model utilizes IGDFS to choose the optimal ICO feature set and FSVM-CIL in the prediction of ICO success to tackle the problem of imbalanced sampling. The system has two major phases: the training and testing phases. The next subsections discuss the system's components in detail, with clarification of the objective of each step.

A. DATASET DESCRIPTION

Two approaches exist for sampling for ICO success prediction: balanced sampling and imbalanced sampling. In this regard, the datasets used in this study are open source and related to two published papers. The first dataset (dataset1) is based on a sample collected and employed by Fahlenbrach and Frattaroli [30]. This dataset is a hand-collected dataset on token sales from primary sources such as whitepapers or other documents published by issuers, archived issuer websites, company announcements on social media, source code on Github, company announcements on bitcointalk.org message boards, and varied national commercial registers. Table 1 defines the attributes selected from this dataset in detail. This sample contains 306 ICOs. However, SMOTE is utilized to increase the number of observations in the minority class (failed ICOs), making the imbalance ratio 6 to 1. The final sample after implementing SMOTE is 351 observations. Features that have more than 30% missing values are discarded. A median filter is applied to handle missing values for the remaining features.

The second dataset (dataset 2) is based on a sample collected and employed by Cerchiello, *et al.* [35]. This dataset comprises 196 ICOs. Information is gathered from web-based sources, mostly rating platforms such as: icobench.com, TokenData.io, ICODrops.com, and CoinDesk.com. Table 1 defines the attributes chosen from this dataset in detail. The imbalance ratio is 3 to 1 (success vs. failed or scam ICOs).

B. FEATURE EXTRACTION

This step represents the main contribution of the research and includes two stages: feature ranking and feature selection. It is based on the features collected from previous studies shown in Table 1.

C. FEATURE RANKING BASED ON INFORMATION GAIN

Firstly, the features are ranked in order of importance to decision making and classification by measuring the information gain. Several ways exist for feature scoring, like information entropy, correlation, chi-squared test, and Gini index [10], [11]. Entropy is one of numerous methods to estimate diversity. The impurity of information can be estimated via information entropy for quantifying the uncertainty of forecasting the value of the goal feature [14]. Let y be a discrete random variable having two possible outcomes. The binary entropy function H, expressed in logarithmic base 2, i.e., Shannon unit, is defined:

$$H\left(y\right) = -p(+)log_2 p\left(+\right) - p(-)log_2 p\left(-\right) \tag{1}$$

where (+, –) denote the classes, p(+) denotes the probability that a sample $y \in (+)$ and $p(-)$ denotes the probability that $y \in (-)$. Entropy gauges the uncertainty of each variable in the procedure of making decisions. The conditional entropy of two events X and Y, when X has a value of x, can be computed as:

$$\begin{aligned} H\left(Y|X\right) &= \sum\nolimits_{x\epsilon X} p_x\left(x\right) H\left(Y\,|\,X = x\right) \\ &= -\sum\nolimits_{x\epsilon X} p_x\left(x\right) \sum\nolimits_{y\epsilon Y} p_x\left(y|x\right) log_2 p_x\left(y\,|\,x\right) \\ &= -\sum\nolimits_{x\epsilon X} \sum\nolimits_{y\epsilon Y} p_{xy}\left(x, y\right) log_2 p_y\left(y\,|\,x\right) \\ &\lim_{x \to 0} x\, log_2\left(x\right) = 0 \end{aligned} \tag{2}$$

The lower the degree of impurity, the higher the skewness of the class distribution. When the class distribution is uniform, the entropy and the misclassification error are greatest. The lowest value of entropy is attained when all the samples are members of the same class. Information Gain (*IG*) is broadly applied to high- dimensional data to assess the effectiveness of variables in classification. It is the predicted volume of information, i.e., a decrease in entropy. Specifically, the *IG* from a variable *x* is denoted by:

$$IG\,(y|x) = H\,(y) - H\,(y\,|\,x) \tag{3}$$

Higher *IG* means better discriminative power for taking decisions. *IG* is a good estimate to decide the relevance of a variable for classification. The significance of variables towards decision-taking in the model is established by evaluating them using the *IG* measurement. Not all the data attributes are generated evenly and not all of them contribute evenly to the decision making. Therefore, it is possible to sort the attributes in the order of their contribution in decision making through listing the variables in a descendent order of *IG* scores.

D. FEATURE SELECTION USING GENETIC ALGORITHM

The suggested model has to extract the best features that optimize classification outcomes and highlight the discrepancy among different classes. The optimization objective is to discover the best likely solution or solutions to a problem, regarding one or more criteria. Therefore, GA is used to pick out the optimal ICO features and decrease the dimensionality of the training dataset. Hence, the findings from the preceding stage are integrated into the information directed feature selection technique through GA.

GA is an adaptive mutation technique that performs a heuristic search, inspired by the evolution process of genetics. A population, consisting of competitive solutions, is preserved. It is subject to selection, crossover, and mutation to evolve and converge to the optimal solution. A parallel search is executed on the solution space to find an optimal solution while not getting stuck in a local optimum. GA can be used to find good features in a high-dimensional space because it can handle both the size of the search space and the distributions of the features [63].

In the current research, IG is implemented to rank the features and then uses the highest ranked features as an intail population for GA procedure for optimal features selection. GA utilizes the LR classifier as a fitness function. Generally, the prerequisites for searching for an optimum solution in the entire feature space involve a search engine with an initial state, a state space, and a termination condition [64]. Given n number of ranked features, the search space size is $2^{\mathrm{n}} - 1$. As every feature has two possible states: "1" or "0," an n bit string shall have 2^{n} possible combinations. Assume τ features that are not significant to decision-taking with regard to the values of their information gains are removed. The length of a binary string turns out to be $n - \tau$. Even in the decreased feature space ($2^{\mathrm{n}-\tau}$), a brute-force search for a big space of $2^{\mathrm{n}-\tau}$ is yet unfeasible. Evidently, that decrease in space is valuable for GA search. The GA components are:

1- Chromosome: GA preserves a diverse population $x_{1\ldots\mathrm{n}} =< x_1, \ldots, x_{\mathrm{n}} >$ of n individuals x_{i}, the candidate solutions. The fitness of these individuals is assessed through calculating an objective function $f(x_{\mathrm{i}})$. These individual solutions are represented as 'chromosomes' that encompass the whole range of possible outcomes. In this research, binary bit string is utilized to represent a chromosome. The bit strings that represent the genotype (abstract representation) must be converted to phenotype (physical make-up), i.e., feature index representation. The number n of bits signifies the number of features. If the i-th bit is 1, the feature x_{i} is chosen in the subset, and if it is 0, the feature x_{i} is not chosen.

2- Selection operator: selection is the procedure of assessing the fitness of individuals and selecting them for reproduction. There are many methods to carry out selection. A number of commonly employed techniques include: tournament selection, roulette-wheel selection, rank selection, hierarchical selection, and elitist selection. The suggested model has utilized tournament selection to pick adequately good individuals for mating. It is efficient to code, works on parallel architectures and allows the selection pressure to be easily adjusted.

3- Crossover operator: a crossover operator produces two offspring from the two chosen parent chromosomes through the interchanging of portions of their genomes. Crossover is the procedure of eliciting the best genes from parents and reconstructing them into possibly superior offspring. The simplest form of crossover is single-point crossover. Other types are two-point crossover, uniform crossover. A single point crossover has been utilized in the current research. Using 1-point crossover means that offspring genomes will be less diverse, they will be quite similar to their parents.

4- Mutation operator: mutation keeps the genetic variation of a population from one generation of chromosomes to the subsequent generation and raises the potential of the algorithm to produce fitter individuals. With minor mutation likelihood, a character at every position in the string is altered randomly. Mutation of bit strings flips the bits with a small probability in random positions. Uniform mutation has been utilized in the current research.

5- Elitism: elitism ensures that the fittest members are transmitted to the subsequent generation. The top individual or a set proportion of the best fit members exists to the subsequent generation. Low elitism in comparison to the size of the population results in a good balance between variation and non-overfitting conditions. Large elitism makes the best-fit individuals predominate in the population, leading to ineffectual search. The current research ensures that two elite offspring subsist into the subsequent generation.

6- Diversity: population diversity is a substantial element affecting the performance of the genetic search. Diversity guarantees that the solution space is appropriately explored, specifically in the earlier stages of the process of optimization. Very small diversity leads to the premature convergence of GA. The initial range of the population and the mutation amount impact the population's diversity. This research has used tournament selection and uniform mutation in the GA evolutionary procedure.

7- Termination criteria: three potential termination criteria might be utilized for the GA: attaining a satisfactory solution, achieving a predetermined maximum number of generations, the convergence of the population to a particular level of genetic diversity [65]. The algorithm convergence is sensitized to the mutation probability: a very large mutation ratio averts the search from convergence, while a very small ratio leads to premature convergence of the search. The maximum number of generations is employed as the termination criterion in the current research.

8- Fitness function: a fitness function assesses the goodness of every individual in the population in every generation compared to the optimization criterion. To create the subsequent generation, the best-fit individuals are permitted to reproduce via the established crossover and mutation rate. In the current research, linear regression (LR) is employed as the induction algorithm for fitness evaluation.

Suppose $g(x)$ is the mapping function of machine learning algorithm. Given an x, the state of the goal feature can be assessed, i.e. $y = g(x)$. Suppose A is the accuracy achieved through the classification algorithm. It can be evaluated using the function: $A = \varphi\ (\hat{Y}, Y)$, where Y is the list of goal states, and $\hat{Y}$ is the list of predicted goal states for the whole test points. As classification accuracy is utilized as the fitness value f, then:

$$f = (g\,(x)\,|_D, Y) \tag{4}$$

where D is the test set. The GA method with LR is indicated as GA-LR. Algorithm 1 gives the operational steps of the suggested approach of Information Gain Directed Feature Selection (IGDFS).

E. CLASSIFICATION USING FSVM-CIL

Classification is the decision-making process that makes use of the features extracted from the prior stage. The classifier is taught with the training data and then tested with the testing data to recognize the different classes. The suggested model utilizes the FSVM-CIL classifier. FSVM-CIL is an SVM classifier that weighs every data point in a different way so as to accurately classify some significant training examples. If FSVM-CIL is implemented on the imbalanced dataset, it can increase the significance of the minority class data through assigning a larger weight for the minority class.

Algorithm 1 Information Gain Directed Feature Selection.

```
Measure Information Gain of individual features from the dataset
Rank the features In the dataset according to their Importance $F = (f_1 > f_2 > f_3, \ldots.)$
Input Top $m$ feature set $F_r$ and class label C
Output $S$
$S \leftarrow null$
Procedure GA
Input: PopSize Ps, GenSize, GenomeLength N,
  ProbMutation $P_m$
Output: The best individual in all generations
Initialize: Population: Ps*N
  Retain $f_1$ from $F_r$
  Ps← random binary chromosomes
  For each chromosome do
    Compute fitness function LR;
  End for
  Repeat
    Select parents $p_1, p_2$ from population based on
      its fitness;
    For all new children do
    retain $f_1$ from $F_r$;
    Crossover $p_1, p_2$;
    Mutate each gene in new child chromosome with
      probability $P_m$;
    End for
    Evaluate fitness of new individuals
    Replace least-fit population with new best
      individuals
  Until Stopping Criteria
End procedure
```

In the current study, we regard the FSVM model, which was initially suggested by the authors in [22], as a solution to the issue of outliers and noise. The FSVM model considered uses fuzzy-membership functions to determine a different misclassification cost for every training pattern. The normal SVM-learning algorithm considers all the training points uniformly, and thus, it may be sensitized to outliers and noise [66], [67].

The FSVM sets different fuzzy-membership values m_i (or weights) for different instances to reflect their significance to their own class, where more significant instances are allotted higher membership values, whereas the less significant ones, like outliers and noise, are allotted lower membership values. The SVM soft-margin optimization problem is reformulated as follows:

$$\begin{aligned} &Min\ \left(\frac{1}{2}w.w + C\sum_{i=1}^{l} m_i\varepsilon_i\right) \\ &s.t.\ \ y_i\,(w.\Phi(x_i) + b) \geq 1 - \varepsilon_i \\ \varepsilon_i\ \ &\geq\ \ 0, i = 1, \ldots\ldots l. \end{aligned} \tag{5}$$

$\Phi(x_i)$ is denote the corresponding feature space vector. The membership m_i of a data point x_i is integrated into the objective function, and thus, a lower m_i could decrease the influence of the parameter ε_i in the objective function, where the corresponding x_i is treated as less important. In another view, if we regard C as the cost allocated for a misclassification, then each instance is allocated a different misclassification cost value m_iC, so that more significant instances are allocated higher costs, while less significant instances are allocated lower costs. Therefore, SVM can locate a more robust hyperplane through maximizing the margin by allowing some misclassification of less-important instances, like outliers and noise. For solving the FSVM optimization problem, Eq.(5) is converted into the subsequent dual Lagrangian [22]:

$$\text{Max } W(\propto) = \sum_{i=1}^{l} \propto_{\mathrm{i}} - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{1} \propto_{\mathrm{i}} \propto_{\mathrm{j}} y_{\mathrm{i}} y_{\mathrm{j}} K(x_{\mathrm{i}}.x_{\mathrm{j}})$$

$$s.t. \sum_{i=1}^{l} y_i \propto_i = 0,\ 0 \leq \propto_i \leq C,\ i = 1, \ldots\ldots..l. \qquad (6)$$

$K(x_{\mathrm{i}}.x_{\mathrm{j}})$ is a function called *kernel* that can compute the dot product of the data points in feature space. The FSVM model has been successfully implemented to decrease the impact of outliers and noise in diverse domains with different means of setting fuzzy-membership values (see [53]–[57]). However, it can still be subject to the issue of class imbalance since there is no modification in the FSVM model in comparison to the original SVM to make it less sensitized to class imbalance. Hence, FSVM-CIL was introduced to overcome the issue of class imbalance. In the suggested FSVM-CIL technique, the membership values for training samples are allocated in such a manner to fulfil the following two objectives: (1) suppressing the influence of class imbalance, and (2) reflecting the within-class significance of different training examples so as to repress the influence of outliers and noise.

Let m_i^+ denotes the membership value of a positive-class example x_i^+, while m_i^- denotes the membership of a negative class example x_i^- in their own classes. In the suggested FSVM-CIL technique, these membership functions are denoted as follows:

$$m_i^+ = f\left(x_i^+\right) r^+ \qquad (7)$$

$$m_i^- = f\left(x_i^-\right) r^- \qquad (8)$$

where $f(x_i)$ outputs a value between 0 and 1, which represents the significance of x_i in its own class. Moreover, the values for r^+ and r^- are allocated so as to represent the class imbalance, such that $r^+ > r^-$, Hence, a positive-class example can take a membership value in the $\left[0, r^+\right]$ interval, whereas a negative-class example can take a membership value in the $\left[0, r^-\right]$ interval. This way of assigning membership values can be used to deal with both class imbalance and outliers/noise at the same time.

IV. EXPERIMENTAL RESULTS

In this part, we first show the chosen features and the experimental set-up. Secondly, for comparison, we implement several imbalanced classification algorithms, including the suggested hybrid IGDFS-FSVM.

A. EXPLANATORY DATA ANALYSIS

The ICO success or failure is utilized as the dependent variable for the imbalanced classification algorithm. Dataset 1 comprises a group of 57 features. Applying GA leads to select15 features, whereas applying IGDFS leads for selecting 10 features. Dataset 2 involves a group of 22 features. Applying GA leads for selecting 8 features, whereas applying IGDFS leads to select 4 features. The selected features are demonstrated in Table 2.

B. EXPERIMENTAL SET-UP

1) PARAMETERS FOR CLASSIFICATION METHODS

For the sake of demonstrating the efficacy of the suggested FSVM-CIL in predicting ICO success, a comparison is made between it and several state-of-the-art algorithms, including Easy Ensemble [58], cost-sensitive adaptive boosting (cs-AdaBoost) [59], cost-sensitive Random Forest (cs-RF) [60], random under-sampling boosting (RUSBoost) [61], and balanced bagging. The basis of the comparison is the prediction accuracy, besides three other measures, precision, and recall. AUC is utilized as a general evaluation criterion for imbalanced classification whereas precision and recall are additional ratios to examine if an imbalanced classification exhibits a decent performance in ICO evaluation [62]. For SVM-based learning machines, the radial basis function kernel is employed. For the FSVM procedure, the gaussian_kernel is used. For FSVM, the regularization parameter C is set at 50, gamma is set at 0.00001, and sigma is set at 3. See [22] for FSVM parameters configuration. The data is split into 80% training and 20% testing. Table 3 is a confusion matrix that visualizes the classification performance.

- *Accuracy* is the ratio of correctly classified firms. It is one of the most extensively employed evaluation metrics.

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

where *TP*, *TN*, *FP*, and *FN* respectively represent true positive, true negative, false positive, and false negative. TP is the number of correctly classified failed ICOs. *TN* is the number of correctly classified successful ICOs. *FP* is the number of successful ICOs misclassified as failed. *FN* is the number of failed ICOs misclassified as successful.

- *Precision Score* is the number of classified failed ICOs which actually failed.

$$Precision = \frac{TP}{TP + FP} \qquad (10)$$

- *Recall Score* gauges how well a classification algorithm can identify failed ICOs. It is also known as the sensitivity metric. A classification algorithm with a greater *TP* ratio is more valuable to investors in

TABLE 2. The selected features set.

Cat.	Features	Data 1 GA	Data 1 IG + GA	Data 2 GA	Data 2 IG + GA
ICO Characteristics	Token_profit	✓	✓		
	Pre-ICO/presale		✓		
	Team tokens locked up		✓		
	Team lockup Period (weighted avg.)		✓		
	KYC	✓	✓		
	Product can be tried out	✓			
Disclosure Character.	Funding milestones	✓			
	Business model available	✓	✓		
	Legal advisor disclosed	✓			
	Whitepaper				✓
Social Media	Twitter			✓	
	Medium				✓
	# Telegram			✓	
	Telegram chat Sentiment Score _ BING			✓	✓
	Positive words_ BING			✓	
	Positive words_ NRC			✓	
	Negative words_ NRC				✓
Financial Details	Token share team (ex ante)	✓			
	USD Price of Token	✓			
	Use of proceeds mentioned	✓			
	ICO Hard Cap	✓			
	Unsold tokens kept by issuer	✓	✓		
Team Characteristics	Team Size	✓	✓	✓	
	Advisors			✓	
	# Advisors			✓	
	Experienced team	✓			
	Team experience missing	✓			
	Financial advisor disclosed		✓		
	Team member with business background	✓	✓		
Total features		15	10	8	4

minimizing their possible investment loss.

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

- *Area under ROC curve* (*AUC*): The AUC of a classification algorithm is equal to the likelihood that the classification algorithm will rank a randomly selected positive example higher than a randomly selected negative example. Then, AUC is employed as a comparison metric for the classification performance of every learning machine [68]. AUC can be delineated as follows.

$$AUC = \frac{(1 + TP_{rate} - FP_{rate})}{2} \tag{12}$$

where TP_{rate} and $\mathrm{FP}_{\mathrm{rate}}$ are the percentage of positive examples that are correctly classified and the percentage of negative examples that are incorrectly classified.

C. ANALYSIS OF EXPERIMENTAL RESULTS

Table 4 demonstrates the confusion matrix of the prediction outcomes of the FSVM-CIL classifier implemented on various feature sets. For dataset 1, applying FSVM-CIL on the entire feature set yields a 11.27% prediction error in the minority class. On the contrary, the application of FSVM-CIL on features selected by GA or features selected by IGDFS results in no prediction errors. For dataset 2, applying FSVM-CIL on the entire feature set yields a 12.5% prediction error on the majority class and a 5% prediction error in the minority class. On the contrary, applying FSVM-CIL on features selected by GA or features selected via IGDFS yields only a 2.5% prediction error in the majority class.

TABLE 3. Confusion matrix.

Total Population		Predicted	
		Successful	Failed
Actual	Successful	True Success	False Failed
	Failed	False Success	True Failed

Table 5 gives full statistics of the performance measures for the suggested hybrid IGDFS-FSVM algorithm in comparison with several state-of-the-art classifiers addressing imbalanced data, SVM-based learning machines, and a hybrid GA-FSVM. For dataset 1, the suggested model is superior to all other algorithms. More specifically, hybrid IGDFS-FSVM and hybrid GA-FSVM attain the best performance (100%) in all measures. Nevertheless, the suggested hybrid IGDFS-FSVM model relies on fewer features (10 vs. 15) in comparison with the hybrid GA-FSVM to attain equivalent performance.

Taking a closer look at Table 5, the accuracy of the eight other algorithms is in the range of 84.5% to 98.59%, and AUC is in the range of 50% to 97.58%. The second best model following the hybrid models with regard to accuracy is RUSBoost, which achieves an accuracy of 98.59% and an AUC of 94.4%. Balanced Bagging attains better outcomes than RUSBoost regarding AUC (97.58%) because of maximum recall and accurate prediction of all observations in the failed class. While Cs-AdaBoost and Cs-RF have almost equal accuracy (97%), Cs-AdaBoost has a higher AUC (93.6% vs. 88.8%) because of poor recall of Cs-RF. In contrast, SVM and Cs-SVM have the poorest performance in all measures, where they act as dummy classifiers and cannot recognize failed ICOs (0% recall), maximizing losses for potential investors.

Similarly, for dataset 2, the suggested model outperforms all other algorithms except for the hybrid GA-FSVM. Significantly, both hybrid IGDFS-FSVM and hybrid GA-FSVM yield accuracy of 97.5%, AUC of 98.48%, precision of 87.5%, and recall of 100%. Nevertheless, the suggested hybrid IGDFS-FSVM model relies on fewer features (4 vs. 8) than the hybrid GA-FSVM. The accuracy of the eight other algorithms is in the range of 82.5% to 92.5%, and AUC is in the range of 78% to 96.96%. More specifically, Balanced Bagging performs second best subsequent to the hybrid algorithms, achieving an accuracy of 95% and an AUC of 96.96%. While Easy Ensemble, Cs-AdaBoost, and

TABLE 4. Confusion matrix for the prediction results of fsvm-cil using various sets of features.

# Features		Dataset 1		Dataset 2	
		Success	Failure	Success	Failure
All Features	Success	87.32%	0	70%	12.5%
	Failure	11.27%	1.41%	5%	12.5%
GA Features	Success	87.32%	0	80%	2.5%
	Failure	0	12.68%	0	17.5%
IGDFS Features	Success	87.32%	0	80%	2.5%
	Failure	0	12.68%	0	17.5%

TABLE 5. Comparison between classifiers and the proposed hybrid igdfs-fsvm.

Classifiers	Dataset 1				Dataset 2			
	Accuracy	AUC	Precision	Recall	Accuracy	AUC	Precision	Recall
EasyEnsemble	95.77%	92.8%	80%	88.89%	92.5%	95.45%	70%	100%
Cs-AdaBoost	97%	93.6%	88.89%	88.89%	92.5%	95.45%	70%	100%
Cs-RF	97.18%	88.8%	100%	77.78%	92.5%	95.45%	70%	100%
RUSBoost	98.59%	94.4%	100%	88.89%	90%	93.9%	63.64%	100%
BalancedBagging	95.77%	97.58%	75%	100%	95%	96.96%	77.78%	100%
SVM	87.3%	50%	0	0	82.5%	89%	50%	100%
Cs-SVM	84.5%	50%	0	0	82.5%	89%	50%	100%
FSVM	88.7%	55%	100%	11.11%	82.5%	78%	50%	71.43%
Hybrid GA-FSVM	100%	100%	100%	100%	97.5%	98.48%	87.5%	100%
Hybrid IGDFS -FSVM	100%	100%	100%	100%	97.5%	98.48%	87.5%	100%

TABLE 6. The improvement ratio of the proposed hybrid igdfs–fsvm compared with other classifiers.

Classifier	Dataset 1				Dataset 2			
	Accuracy	AUC	Precision	Recall	Accuracy	AUC	Precision	Recall
EasyEnsemble	4.23%	7.2%	20%	11.11%	5%	3.03%	17.5%	0
Cs-AdaBoost	3%	6.4%	11.11%	11.11%	5%	3.03%	17.5%	0
Cs-RF	2.82%	11.2%	0	22.22%	5%	3.03%	17.5%	0
RUSBoost	1.41%	5.6%	0	11.11%	7.5%	4.58%	23.86%	0
BalancedBagging	4.23%	2.42%	25%	0	2.5%	1.52%	9.72%	0
SVM	12.7%	50%	100%	100%	15%	9.48%	37.5%	0
Cs-SVM	15.5%	50%	100%	100%	15%	9.48%	37.5%	0
FSVM	11.3%	45%	0	88.89%	15%	20.48%	37.5%	28.57%

Cs-RF have equal accuracy of (92.5%) and AUC of (95.45%). These models have zero misclassifications in the minority class (failed class), achieving 100% recall but only 70% precision. In contrast, SVM, Cs-SVM, as well as FSVM have the poorest performance among all models.

Furthermore, the superiority of the suggested model is demonstrated in Table 6, which outlines the decreased error rates when comparing the suggested model with other classifiers. For dataset 1, the accuracy improved by about 1.41% to 15.5%, and the AUC was improved by about 2.42% to 50%. For precision score, the proposed model obtained higher results than other classifiers except for Cs-RF, RUSBoost, and FSVM-CIL. For recall score, the suggested model obtained superior results than all algorithms except Balanced Bagging.

Similarly, for dataset 2, the accuracy improved by about 2.5% to 15%, and the AUC improved by about 1.52% to 20.48%. For precision score, the suggested model outperformed all other classifiers. For recall score, the suggested model obtained superior results than FSVM-CIL. The results indicate that the suggested model more precisely predicts observations in the majority class (success) in comparison with other algorithms. Collectively, it can be concluded that the suggested hybrid model is a plausible and robust classification algorithm with the highest performance on all measures.

V. CONCLUSION

This study presents a hybrid intelligent model for forecasting the success/failure of ICOs. We first apply IGDFS to select the optimum features, and thereafter, implement the FSVM-CIL classifier to differentiate failed and successful ICOs in two imbalanced datasets. The use of information gain to direct GA in feature selection reduces the number of features chosen while maintaining the same level of performance.

To demonstrate the efficacy of the suggested IGDFS-FSVM system, we compared it with several imbalanced classifier benchmarks. The classification results reveal that the suggested IGDFS-FSVM outperforms other classifiers. The performance of IGDFS-FSVM is equal to that of GA-FSVM in all evaluation measures. However, IGDFS-FSVM selects fewer features from both datasets. Considering that the ICO success prediction issue in the FinTech market separates the minority class from the majority class, we infer that the utilization of IGDFS-FSVM is successful since it improves the classification performance and well detects failed ICOs. Based on empirical findings, we draw the conclusion that the hybrid IGDFS-FSVM model can be utilized as a decent classifier for ICO investment decisions. Future work includes utilizing other types of optimization for features selection process.

REFERENCES

[1] M. Iansiti and K. Lakhani, "R. (2017). The truth about blockchain," *Harvard Bus. Rev.*, vol. 95, no. 1, pp. 19–127.

[2] P. P. Momtaz, "Initial coin offerings," *PLoS ONE*, vol. 15, no. 5, 2020, Art. no. e0233018.

[3] P. P. Momtaz, "Token sales and initial coin offerings: Introduction," *J. Alternative Investments*, vol. 21, no. 4, pp. 7–12, Mar. 2019.

[4] C. Catalini and J. S. Gans, *Initial Coin Offerings and the Value of Crypto Tokens*. Cambridge, MA, USA: National Bureau of Economic Research, 2018.

[5] J. Li and W. Mann. (2018). *Initial Coin Offerings and Platform Building*. [Online]. Available: https://ssrn.com/abstract=3088726

[6] P. P. Momtaz, "Entrepreneurial finance and moral hazard: Evidence from token offerings," *J. Bus. Venturing*, vol. 36, no. 5, pp. 1–24.

[7] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akinade, and M. Bilal, "Systematic review of bankruptcy prediction models: Towards a framework for tool selection," *Expert Syst. Appl.*, vol. 94, pp. 164–184, Mar. 2018.

[8] C.-H. Chou, S.-C. Hsieh, and C.-J. Qiu, "Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction," *Appl. Soft Comput.*, vol. 56, pp. 298–316, Jul. 2017.

[9] J. Sun, H. Li, Q.-H. Huang, and K.-Y. He, "Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches," *Knowl.-Based Syst.*, vol. 57, pp. 41–56, Feb. 2014.

[10] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

[11] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, May 2003.

[13] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.

[14] S. Jadhav, H. He, and K. Jenkins, "Information gain directed genetic algorithm wrapper feature selection for credit rating," *Appl. Soft Comput.*, vol. 69, pp. 541–553, Aug. 2018.

[15] J. An, T. Duan, W. Hou, and X. Xu, "Initial coin offerings and entrepreneurial finance: The role of founders' characteristics," *J. Alternative Investments*, vol. 21, no. 4, pp. 26–40, Mar. 2019.

[16] T. Bourveau, E. T. De George, A. Ellahie, and D. Macciocchi. (2018). *Initial Coin Offerings: Early Evidence on the Role of Disclosure in the Unregulated Crypto Market*. [Online]. Available: https://ssrn.com/abstract=3193392

[17] S. T. Howell, M. Niessner, and D. Yermack, "Initial coin offerings: Financing growth with cryptocurrency token sales," *Rev. Financial Stud.*, vol. 33, no. 9, pp. 3925–3974, Sep. 2020.

[18] S. Samieifar and D. G. Baur, "Read me if you can! An analysis of ICO white papers," *Finance Res. Lett.*, vol. 38, Jan. 2021, Art. no. 101427.

[19] J. Tian, H. Gu, and W. Liu, "Imbalanced classification using support vector machine ensemble," *Neural Comput. Appl.*, vol. 20, no. 2, pp. 203–209, 2011.

[20] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.

[21] K. Yoon and S. Kwek, "A data reduction approach for resolving the imbalanced data issue in functional genomics," *Neural Comput. Appl.*, vol. 16, no. 3, pp. 295–306, May 2007.

[22] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.

[23] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy support vector machines for class imbalance learning," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 3, pp. 558–571, Jun. 2010.

[24] E. Mollick and A. Robb, "Democratizing innovation and capital access: The role of crowdfunding," *California Manage. Rev.*, vol. 58, no. 2, pp. 72–87, Feb. 2016.

[25] V.-S. Ha and H.-N. Nguyen, "FRFE: Fast recursive feature elimination for credit scoring," in *Proc. Int. Conf. Nature Comput. Commun.*, Cham, Switzerland: Springer, 2016, pp. 133–142.

[26] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*, vol. 453. Cham, Switzerland: Springer, 1998.

[27] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, pp. 245–271, Dec. 1997.

[28] D. Koller and M. Sahami, *Toward Optimal Feature Selection*. Stanford, CA, USA: Stanford InfoLab, 1996.

[29] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowl.-Based Syst.*, vol. 86, pp. 33–45, Sep. 2015.

[30] R. Fahlenbrach and M. Frattaroli, "ICO investors," *Financial Markets Portfolio Manage.*, vol. 35, no. 1, pp. 1–59, Mar. 2021.

[31] S. Adhami, G. Giudici, and S. Martinazzi, "Why do businesses go crypto? An empirical analysis of initial coin offerings," *J. Econ. Bus.*, vol. 100, pp. 64–75, Nov. 2018.

[32] S. Bian, Z. Deng, F. Li, W. Monroe, P. Shi, Z. Sun, W. Wu, S. Wang, W. Y. Wang, A. Yuan, T. Zhang, and J. Li, "IcoRating: A deep-learning system for scam ICO identification," 2018, *arXiv:1803.03670*.

[33] Charlotte, S.-H. Wu, H.-C. Sung, and T.-C. Cheng, "Measuring ICO performance indicators: An empirical study via white papers," in *Proc. IEEE Int. Conf. Blockchain Cryptocurrency (ICBC)*, May 2019, pp. 128–132.

[34] C. Fisch, "Initial coin offerings (ICOs) to finance new ventures," *J. Bus. Venturing*, vol. 34, no. 1, pp. 1–22, Jan. 2019.

[35] P. Cerchiello, P. Tasca, and A. M. Toma, "ICO success drivers: A textual and statistical analysis," *J. Alternative Investments*, vol. 21, no. 4, pp. 13–25, Mar. 2019.

[36] R. Amsden and D. Schweizer, "Are blockchain crowdsales the new 'gold rush'? Success determinants of initial coin offerings," Success Determinants Initial Coin Offerings, Social Sci. Res. Netw. (SSRN), Rochester, New York, NY, USA, Tech. Rep., pp. 1–66.

[37] S. Albrecht, B. Lutz, and D. Neumann, "The behavior of blockchain ventures on Twitter as a determinant for funding success," *Electron. Markets*, vol. 30, no. 2, pp. 241–257, 2020.

[38] C. Fisch and P. P. Momtaz, "Venture capital and the performance of blockchain technology-based firms: Evidence from initial coin offerings (ICOs)," Social Sci. Res. Netw. (SSRN), Rochester, New York, NY, USA, Tech. Rep., pp. 1–45.

[39] D. Florysiak and A. Schandlbauer, "The information content of ICO white papers," Social Sci. Res. Netw. (SSRN), New York, NY, USA, Tech. Rep., pp. 1–45.

[40] D. Blaseg, "Dynamics of voluntary disclosure in the unregulated market for initial coin offerings," Social Sci. Res. Netw. (SSRN), New York, NY, USA, Tech. Rep., 2021, pp. 1–45.

[41] D. Boreiko and G. Vidusso, "New blockchain intermediaries: Do ICO rating websites do their job well?" *J. Alternative Investments*, vol. 21, no. 4, pp. 67–79, Mar. 2019.
[42] C. Feng, N. Li, M. Wong, and M. Zhang, "Initial coin offerings, blockchain technology, and white paper disclosures," Mingyue, Initial Coin Offerings, Blockchain Technol., Social Sci. Res. Netw. (SSRN), New York, NY, USA, Tech. Rep., 2021, pp. 1–45.
[43] C. Perez, K. Sokolova, and M. Konate, "Digital social capital and performance of initial coin offerings," *Technological Forecasting Social Change*, vol. 152, Mar. 2020, Art. no. 119888.
[44] J. Sun, Z. Shang, and H. Li, "Imbalance-oriented SVM methods for financial distress prediction: A comparative study among the new SB-SVM-ensemble method and traditional methods," *J. Oper. Res. Soc.*, vol. 65, no. 12, pp. 1905–1919, 2014.
[45] V. Vapnik, *The Nature of Statistical Learning Theory*. Cham, Switzerland: Springer, 1999.
[46] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Apr. 1995.
[47] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. (AI)*, vol. 55, 1999, p. 60.
[48] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," *Mach. Learn.*, vol. 2004, pp. 39–50, Mar. 2004.
[49] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *Proc. ICML Workshop Learn. From Imbalanced Data Sets II*. Washington, DC, USA: Citeseer, 2003, pp. 49–56.
[50] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[52] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 6, pp. 820–831, Dec. 2005.
[53] H.-L. Dai, "Class imbalance learning via a fuzzy total margin based support vector machine," *Appl. Soft Comput.*, vol. 31, pp. 172–184, Jun. 2015.
[54] W. Lee, C.-H. Jun, and J.-S. Lee, "Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification," *Inf. Sci.*, vol. 381, pp. 92–103, Mar. 2017.
[55] S. Ando, "Classifying imbalanced data in distance-based feature space," *Knowl. Inf. Syst.*, vol. 46, no. 3, pp. 707–730, 2016.
[56] Q. Fan, Z. Wang, D. Li, D. Gao, and H. Zha, "Entropy-based fuzzy support vector machine for imbalanced datasets," *Knowl.-Based Syst.*, vol. 115, pp. 87–99, Jan. 2017.
[57] P. Cho, M. Lee, and W. Chang, "Instance-based entropy fuzzy support vector machine for imbalanced data," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1183–1202, 2019.
[58] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[59] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, vol. 96, 1996, pp. 148–156.
[60] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[61] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.
[62] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Mach. Learn.*, vol. 31, no. 1, pp. 1–38, 2004.
[63] L. Li, W. Jiang, X. Li, K. L. Moser, Z. Guo, L. Du, Q. Wang, E. J. Topol, Q. Wang, and S. Rao, "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16–23, Jan. 2005.
[64] R. Kohavi and G. H. John, "The wrapper approach," in *Feature Extraction, Construction Selection*. Cham, Switzerland: Springer, 1998, pp. 33–50.
[65] M. M. Lankhorst, *Genetic Algorithms in Data Analysis*. Groningen, The Netherlands: Rijksuniversiteit Groningen, 1996.
[66] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory (COLT)*, 1992, pp. 144–152.
[67] X. Zhang, "Using class-center vectors to build support vector machines," in *Proc. Neural Netw. Signal Process. IX IEEE Signal Process. Soc. Workshop*, Aug. 1999, pp. 3–11.
[68] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.

MOHAMED GIHAN ALI received the B.Sc. degree in accounting from the Faculty of Commerce, Alexandria University, Damanhour Branch, Egypt, in 2008, the M.Sc. degree in accounting from the Accounting Department, Faculty of Commerce, Alexandria University, in 2013, and the Ph.D. degree from Alexandria University for a thesis in corporate bankruptcy prediction based on statistical and machine learning techniques, where she is currently pursuing the M.Sc. degree in information technology with the Institute of Graduate Studies and Research. She is the author or coauthor of some articles publications in the field of accounting, auditing, and machine learning. She completed many online courses and projects in data science, machine learning, and Fintech from highly esteemed universities, such as Stanford, Michigan, and Pennsylvania, besides the Advanced Data Analysis Nanodegree from Udacity. Her research interests include financial accounting, managerial accounting, auditing, financial inclusion, data analytics, and machine learning.

ISMAIL IBRAHIM GOMAA received the Ph.D. degree in accounting from the University of Florida. He had positions as the Chairperson of the Accounting Department and the Department of Information and Computer Systems, Alexandria University. He worked as a Faculty Member at the University of Florida and The Ohio State University. He is currently a Professor of accounting and information systems, and the Former Dean of the Faculty of Commerce, Alexandria University. He has been recognized by Alexandria University for pioneer work and outstanding academic contribution. His research interests include financial accounting, auditing, accounting information systems, forensic accounting, and optimization techniques. He is a member of the editorial board of five international journals and served as a reviewer for research published in highly ranked journals and international conferences. In addition to his extensive publications, he has supervised 57 Ph.D. dissertations and master's theses and developed undergraduate and graduate programs in accounting and information systems for seven universities in Egypt and the region. He is a member of several academic and professional organizations and has extensive experience as a Financial Consultant. He is also the Chairperson of the Board of Directors of one of the largest corporations in Egypt.

SAAD MOHAMED DARWISH received the B.Sc. degree in statistics and computer science from the Faculty of Science, Alexandria University, Egypt, in 1995, the M.Sc. degree in information technology from the Department of Information Technology, Institute of Graduate Studies and Research (IGSR), University of Alexandria, in 2002, and the Ph.D. degree from Alexandria University for a thesis in image mining and image description technologies. Since June 2017, he has been a Professor with the Department of Information Technology, IGSR. He is the author or coauthor of more than 50 papers publications in prestigious journals and top international conferences and also received several citations. He has supervised around 60 M.Sc. and Ph.D. students. His research interests include image processing, optimization techniques, security technologies, database management, machine learning, biometrics, digital forensics, and bioinformatics. He has served as a reviewer for several international journals and conferences.