National College of Ireland

Detecting Anomalous Insurance Claims with Hybrid Feature Optimisation and Classification Techniques

MSc Research Project
FinTech

Sananda Dasgupta
Student ID: X18115781

School of Computing
National College of Ireland

Supervisor: Mr. Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sananda Dasgupta
Student ID:	X18115781
Programme:	MSc. in FinTech **Year:** 2018-19
Module:	Research Project
Supervisor:	Victor Del Rosal
Submission Due Date:	12th August 2019
Project Title:	"DETECTING ANOMALOUS INSURANCE CLAIMS WITH HYBRID FEATURE OPTIMISATION AND CLASSIFICATION TECHNIQUES"
Word Count:	6692 **Page Count** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

DETECTING ANOMALOUS INSURANCE CLAIMS WITH HYBRID FEATURE OPTIMISATION AND CLASSIFICATION TECHNIQUES

Sananda Dasgupta
X18115781

Abstract

As the world is gradually being engulfed by the inevitability of technology, each and every aspect that technology is incorporated into are becoming increasingly vulnerable to digital crimes. The insurance sector is no exception to this and the potentiality of insurance frauds are taking a huge toll on the industry and the numbers are increasing day by day. Basically, any act carried out to swindle an insurance process can be termed as an 'Insurance Fraud'. Fake claims are another way by which a malefactor can deceive an insurance process. Such is the magnitude of the threat that the industry loses almost $30 billion a year according to a recent survey. Several methods and processes have been applied and tested as an anti-fraud measure to minimise and ideally terminate illicit activities in the sector and data-mining methodologies have proven to be instrumental in fighting digital crimes in the insurance domain. Although there exists several ways and methods of applying data-mining into a fraud-prevention program, this research particularly aims to explore an optimal hybrid model in identification of aberrant and atypical activities in an insurance claim process in an attempt to detect potential anomaly. The efficiency of this particular model that combines feature optimisation with classification algorithms is based on the performance metrices viz. accuracy, sensitivity and specificity. The model is being tested on a dataset of insurance claim taken from Kaggle and the feature optimisation algorithms used are Particle Swarm Optimisation (PSO) and Firefly Algorithm (FFA). The classification algorithms applied are Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes (NB), K-Nearest Neighbour (kNN) and Random Forest (RF). In an attempt to achieve a high quality predictive output on the basis of the above-mentioned metrices, this paper investigates that a hybrid combination of PSO and RF proves to be the most effective in achieving the best predictions over other models. This research is optimistic that the model deduced will hence allow insurers to put a check on fraudulent activities within the industry that eventually will save billions of dollars globally.

Keywords: Data Mining, Feature Selection, Fraudulent Insurance Claims, Particle Swarm Optimisation (PSO), Firefly Algorithm (FFA), Classification Algorithm, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbour (kNN).

1 Introduction

One of the most alarming concerns that haunt the insurance companies is the increase in insurance fraud in the form of fraudulent claims. Forged or fake claims have always been a point of concern for the insurers and this form of crime have proven to be unsettling for many years. As companies evolve technologically, they formulate new methods to put a check to the problem but on the other side of the spectrum individuals with harmful intentions continue to evolve simultaneously, finding loopholes into the system to perform their illicit activities (Warren *et al.*, 2018). As per data, insurance frauds are recorded mostly in the form

of stolen car or accident scams, false health insurance billing to even a faked death. In spite of the fact that the fraud detection system is heavily reliant on several systems based on pattern matching (Verma *et al.*, 2017), it proves to be ineffective in most cases, as the associated risks often appear in clusters. Simply put, the existence of one risk factor reflects that other risk factors also exist. It can be said that with increasing risk factors the chances of fraudulent incidents also increase. According to "The Association of British Insurers" the frequency of falsified claims is as high as 2500 per week. According to a relatively new study by the "European Healthcare Fraud and Corruption Network" (EHFCN) in association with the "Centre for Counter Fraud Services" (CCFS) conducted at Britain's Portsmouth University, it was derived that error or corrupt work leads to a loss of almost 5.59% of the annual spending on global health. Moreover, it states that health care providers incur loss up to U.S. $260 billion (180 billion Euros) each year that is almost 6% of global health care spending over fraudulent activities.

Data mining is state of the art technique readily used by todays IT professionals to identify or deduce patterns in relatively large datasets. Data mining provides effective tools for studying large datasets to deduce some logical meaning and to extricate patterns, information, and connections that may be extremely difficult and time consuming to decipher with conventional statistical methods (Ngai *et al.*, 2011). Data mining plays an instrumental role in the insurance sector, where classification categorizes a transaction as fraud or genuine, based on their similarities to preceding transaction details. Hence, classification techniques split up group of even instances and assign them to many unique and extensive categories known as classes. This signifies that every object must be assigned to particular class, i.e. a transaction can either be suspicious or legal (Bramer; 2013; B.N.Lakshmi and G.H.Raghunandhan; 2011). The term 'classifier' signifies the applied function by a classification algorithm which maps input data into different classes and extract predictions that are based on historic data (Ghorbani and Ghousi; 2019). Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Trees are some examples of classifiers. The success of the classifier's performance is greatly dependent on feature optimisation algorithms which eliminates inapposite and irrelevant features to derive an accurate and enhanced predictive model (Zhang *et al.*, 2019). Features such as 'policy number' or 'policy state' might not be a factor contributing to risk and is of negligible importance in detecting unlawful claims. In a situation with considerably higher number of attributes, the size of the dataset will be bigger, and the data will be uncleaned and hence the performance of the classifier would be negatively affected (Wahid and Rao; 2019). Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), Firefly Algorithm (FFA) etc are some of the examples of feature selection methods. Combining feature optimisation and classification algorithms will result in a "combinatorial optimisation" which enhances accuracy, sensitivity and specificity in many folds (Liu and Yu; 2005). Here, accuracy signifies the percentage of instances which are classified correctly. Sensitivity is the measure of correct classification of positive instances and specificity denotes the correct classification of the negative instances (Shwartz and David; 2014).

The current study aims to explore, scrutinise and then determine the optimal hybrid model of feature optimisation and classification that identifies malicious insurance claims judged on three performance metrices viz. accuracy, sensitivity and specificity. The implementation of the research lies in the collective application of feature optimisation and classification techniques that is yet to be experimented in detecting fraud within the insurance domain.

However, this study extends only to the use of insurance fraud dataset extracted from Kaggle[1] as it is available for use by the machine learning community for verifiable analysis. Thus, the two-fold goal of this study are as follows:

1. To determine the effectiveness of popularly used feature selection techniques (Particle Swarm Optimisation and Firefly Optimisation) to avoid the cumbersome dimensionality while identifying an anomalous claim.

2. To assess if hybrid machine learning (ML) models combined with feature optimisation methods are useful and effective in prediction of a suspicious claim.

Research Question

"To what extent can a hybrid model of feature optimisation and classification algorithms provide a significant improvement in the detection of fraudulent insurance claims when compared with the state of the art?"

This paper is organized as follows:

- Section 2 brings forth a review of literature which demonstrates a comparative and critical analysis of works that already exist. This work serves as an important reference for the basis of the research. Section 2 also serves as a comparison of tools of data mining.
- Section 3 & 4 demonstrates the research methodology and design specification respectively.
- Section 5 refers to the implementation followed by evaluation and a detailed discussion of results drafted in Section 6.
- Section 7 is the study that concludes the limitations of the proposed work and outlines the future work of the research.

2 Related Work

Section 2.1 presents a thorough analysis on data mining algorithms (Webster and Watson; 2002) and its application in detecting fraudulent insurance claims succeeded by a discussion on data mining techniques in Section 2.2 and lastly, summarises the insights gained as the base of the research for better understanding.

2.1 Detecting Suspicious Insurance Claims using Data Mining Techniques

Depending on the nature and the extent of the fraud, most insurance fraud cases are either classified as a felony or a misdemeanor that is usually carried out by manipulating information and forging documents with the intention of unlawfully benefiting from insurance protection. It can be a false case that is forged and submitted to the insurance company to extract money or by over estimating the damages caused in a real accident. (Camarda *et al.*, 2018). In their global insurance trend analysis report of 2018, EY had stated that insurers were in a unique position to take advantage of multiple data sources for building formidable relationships with customers and also to achieve maximum efficiency. The

[1] https://www.kaggle.com/mervynakash/insurance-claim

findings posit a far-reaching hazard to the insurers as they lose approximately $30 billion a year due to fake claims (EY, 2018)[2].
When compared to complicated physical investigations, an automated data mining process assists insurance companies in making almost immaculate detections of fraud easily (Sönmez *et al.* 2018). Data mining is readily used by experts and professionals in an attempt to infer concealed patterns in large datasets, which might otherwise have been very hard to find using traditional statistics (Li *et al.,* 2017). Evidently, it has proved to be a winning tool for the insurers to identify patterns in an extensive number of insurance claim data (Kirildog *et al.,* 2012). Feature selection is the process of filtering and reducing the inputs for processing and analysis and hence feature selection can be looked into as an optimisation problem whose objective is to find relevant and significant information. Variable selection returns a subset of features and often used in domains where the attributes are higher as compared to the number of data points (Huan Liu and Zhao; 2010). Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), Artificial Bee Colony (ABC), Firefly Algorithm (FFA) are a few examples of feature optimisation and are also known as metaheuristic algorithms. Algorithms that contain stochastic components were often considered as heuristic in the past. However, recent literature refers to them as meta-heuristics. Heuristic means to obtain by trial and error method and meta signifies something of higher level or beyond (Jamjala *et al.;* 2018).
Metaheuristics ideally produce better results than simple heuristics. Inarguably one of the most popular and widely accepted evolutionary algorithm with a wide range of applications is the Genetic Algorithms (GAs), developed by John Holland. Based on Darwin's Natural Selection theory, Genetic Algorithms have proven worthwhile in solving various optimisation problems. In GAs, a plethora of probable solutions for a particular problem are present. The solutions are treated with recombination and mutation processes generating new off springs and the process is again duplicated through generations (similar to natural genetics). Every candidate solution is allocated a fitness value (on the basis of its performance in answering a particular problem) and the fitter solutions are allowed a better opportunity to yield even "fitter" solutions. Benchaji *et al.,* (2019) showed the effectiveness of GA to select the optimal variables while applying on a fraud transaction dataset. Arora *et al.,* (2017), on the other hand, presented another feature optimisation algorithm that was created by Kennedy and Eberhart in 1995, called 'Particle Swarm Optimisation' (PSO). This mechanism is influenced by the act of flocking by animals such as birds, insects and fish, in a uniform formation in their search for fodder. When birds start looking for food, they are not aware of the source, hence they move in random fashion. To some degree, each bird is naturally attracted to the location that is more probable to have food. Once they have discovered such area, they unite there and if they find that area unsuitable to their needs they start to look elsewhere. This study is analogous to this phenomenon where particles are represented by the birds and the solution is represented by the food. PSO uses a similar mechanism to solve an optimisation problem and explores the space of an objective function by modifying the trajectories of individual particles. Each particle traces a piecewise path which can be modelled as a time-dependent positional vector (Tayal *et al.,* 2016). The current study also demonstrates the efficacy of PSO in selection of the most useful features in a dataset. Barrera *et al.*, (2014) promulgated GA is not as effective computationally as PSO. However, a major hiccup was the inability to gauge their effectiveness with real life instances which would be more advantageous. This knowledge, hence, led to the usage of PSO in this paper as its feature optimisation technique.

Inspired by the behavioral pattern of fireflies, Xin-She Yang developed the Firefly Algorithm (FFA) in 2008 which is essentially a metaheuristic algorithm. Fireflies supposedly use their

[2] https://www.ey.com/Publication/vwLUAssets/ey-global-insurance-trends- analysis-2018/$File/ey-global-insurance-trends-analysis-2018.pdf

flashing mechanism to signal and attract other fireflies in their vicinity. Three basic assumptions form the basis of the structure of the algorithm viz. unisexual behavior is exhibited by all fireflies; more the attraction, brighter the light; and that these fireflies' movements are random because all the fireflies are equally bright (Adaniya *et al.*, 2012). According to Erdinc; (2017), the algorithm engages three parameters, viz. attractiveness, randomisation and absorption and it can be safely conjectured that the FFA has a higher success rate when it comes to handling multimodal functions compared to PSO or GA.
Feature optimisation is an effectual method for eliminating unnecessary components in a data, however, the importance of the precision of the generated subset should not be ignored and hence it's essential to apply classification tools for validating and investigating the contributing processes that detects suspicious insurance claims (Herland *et al.*, 2017). In a typical data mining procedure, classification techniques allocate instances in a dataset in order to aim at classes and categories. The primary goal of this is to predict the response variable with precision for every instance that occurs in the data. For instance, a classification model is instrumental in recognizing a claim application that is forged (Priya and Pushpa; 2017). Artificial Neural Network (ANN), Support Vector Machines (SVM), Naïve Bayes (NB), K- Nearest Neighbour (kNN), Random Forest (RF) etc. are few examples of such classifiers.

Many researchers in the past have banked upon the predictive potential of ANN over other statistical models in the field of insurance due to its ability to learn from examples. As Larnyo *et al.*, (2018) brought to light, ANN has the capability to infer concealed dependencies those are not linear, even when noise is present. In spite of the results demonstrating 98.98% accuracy, the study failed to subsume factors such as information on patients, doctors and health care service providers or some other data those could possibly be closely or distantly related. A subsequent study by Jan; (2018) employed ANN to detect fraudulent activities in an enterprise's financial statement for sustainable growth of enterprises and financial markets and successfully returned the best classification result with 90.83% accuracy. However, to ensure the global applicability of the model created in this study, various metrices need to be modified. These metrices depend on the country or region's economic system, it's laws in regard to fraud and financial market practices. According to the authors, the shortcoming of this research also lies in the data selected, that only covers a very small segment of Taiwan and overlooks the external effects of the larger domain.

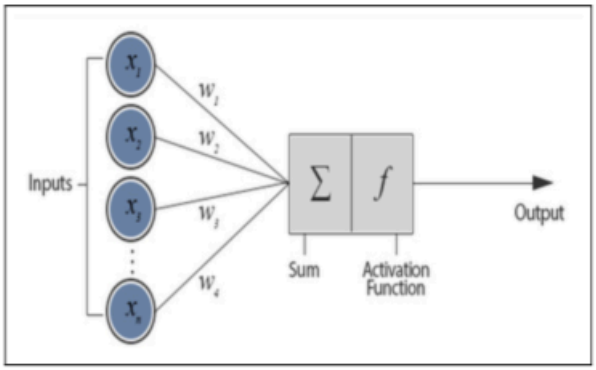

Figure 1: Artificial Neural Network Perceptron

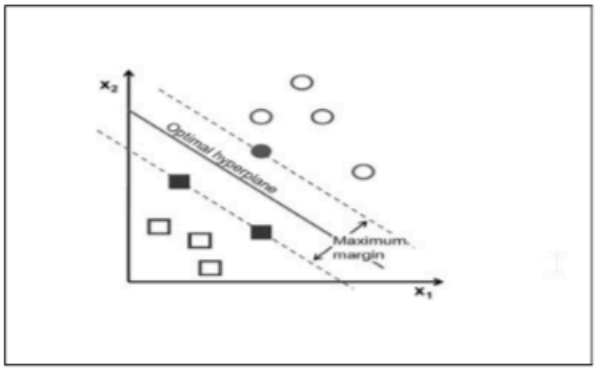

Figure 2: Support Vector Machines (SVM)

Subudhi *et al.*, (2016) applied Support Vector Machine (SVM) alongside fuzzy clustering in his attempt to detect fraud. This particular research affirmed that SVM can be efficacious in detecting fraudulent activities, maintaining the rate of false alarm at its bare minimum and also indicated that by reducing the dimension, performance of the classifier can be radically

improved. Nian *et al.*, (2016) through their work, demonstrated that kNN obtains more accurate results than neural network, although this classifier was not implemented on a real-time instance. Hence, this research digs deep into the performance of kNN combined with diverse feature optimisation mechanisms applied to real-time scenario. According to Viaene *et al.*, (2017), Naïve Bayes (NB) can also be deemed effective as a classifier in forgery identification within the insurance industry. Based on Bayes' theorem, NB can be effective even on a smaller size of data for classification (Rish; 2001). However, this study fell short in scrutinising or exploring the performance of NB when integrated with feature optimisation and the current study aims to explore exactly that loophole. With the intention of assisting stakeholder, Yao *et al.*, (2018) created an optimized decision support model which unveiled Random Forest (RF) to be the optimal performing model among all other four classifiers. While examining the performance of RF incorporated with GA in fraudulent insurance claim detection, West *et al.*, (2017) denoted how 'data preprocessing' has an indispensable role to play in data mining and also demonstrated that the presence of numerous missing values could influence how the classifier performs. Hence, considering the previous results, this research aims to test the effectiveness and the competence of RF combined with PSO and FA respectively.

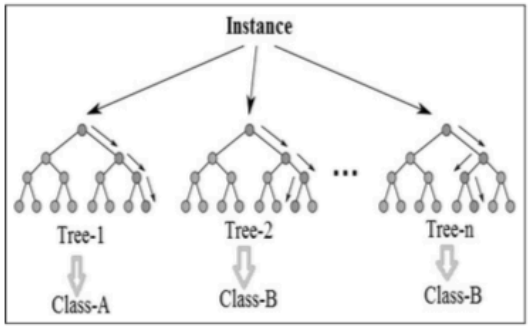

Figure 3: Random Forest

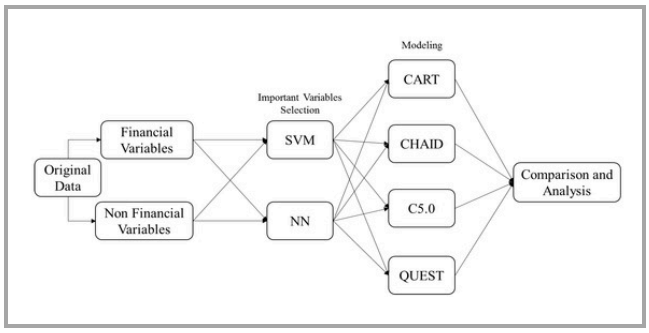

Figure 4: Hybrid Model

Research by Tyagi *et al.*, (2018) demonstrated that hybrid models are not only precise but also more accurate. Therefore, the current research considers this approach by implementing both feature optimisation and classification techniques as its framework for detection of malicious claims in the insurance domain.

Another study by Jan; (2018), successfully performed a new research by using hybrid data mining techniques to detect fraudulent financial statements and illustrated how the hybrid of ANN and CART outperforms the result of other classifiers alone.
Kose *et al.*, (2015) collated various algorithms like Genetic Algorithm, PSO and Neural Network for detection of suspicious claims in healthcare and they have shelved the implementation of pairwise comparison of various optimisation and classification methods for their future work. Nian *et al.*, (2016) propounded a spectral ranking method in the auto-insurance industry for anomaly detection and achieved an accuracy of 74.1%. However, this research also did not take into account any challenging aspects of a real-life dataset as they were expensive as well as time consuming. Sundarkumar *et al.*, (2015) on the other hand, conducted a novel hybrid approach in order to rectify the complication of data imbalance by applying k Reverse Nearest Neighbourhood and One Class Support Vector Machine (OCSVM) simultaneously to detect fraud in automobile insurance claims. In order to ascertain the optimal properties or attributes of a dataset, the researcher has suggested the inclusion of visual symbols and analytical machinery in the future work section of his paper. This could ultimately enhance the classifier performance.

Table 1: Data Mining Algorithms Experimented on Different Fraud Datasets

Author	Year	Technique	Accuracy	Sensitivity	Specificity
Bhattacharyya *et al.*	2010	ANN tuned by genetic algorithm	73.8%	70.4%	65.2%
Wong *et al.*	2012	Artificial Immune System (AIS)	71.3%	65.7%	67.1%
Seeja *et al.*	2014	SVM	55.2%	60.7%	54.5%
		Random Forest	80%	78.2%	72%
		Naïve Bayes	60.4%	62.7%	53.7%
Umarani *et al.*	2015	Optimal Ensemble Classification with PSO (OEC-PSO)	82.6%	71%	77.3%
		Optimal Ensemble Architecture Selection using PSO (OEAS-PSO)	86.3%	78.9%	82.1%
		Optimal Ensemble Architecture Selection using firefly approach (OEAS-FFA)	81.6%	85.2%	77.8%
Yee *et al.*	2018	K2	41.8%	31.3%	39.2%
		Tree Augmented Naïve Bayes	84.2%	75.6%	80.5%
Lee *et al.*	2018	Logistic Clustering (Density Based)	63.1%	68%	62.9%
Jan	2018	ANN+CART	90.83%	85.5%	73.2%
		SVM+CART	85.98%	70.1%	69.6%

Summarising, a general evaluation of the preceding studies that is pertinent to this research in varying standards is depicted in Table 1. The deficiencies of the existing studies are:

- The time required for computation is high.
- The research considers only a particular single fraud type.
- No research till date has attempted to cover the association between all the data and factors.
- Information derived from the model results and its application in practical life are very low.

The result is that no research that were previously conducted, had tried to apply a hybrid model of feature optimisation (PSO/Firefly) and classification to detect fraud in the insurance sector holistically. The fact that there is a dearth of commercially used intrusion detection system in the field of insurance only sustains this conclusion.

2.2 Data Mining Tools

Although there are several researches reviewing data mining algorithms and methods in general but an extensive and comprehensive study on data mining tools are still lacking. (Marjia Sultana and ShorifUddin; 2016; D. P. Shukla and Sen; 2014; Mohammed Abdul Khaled and Dash; 2013). X. Chen and Williams (2007) for example, have analyzed various facets of 12 open source data mining tool against features such as "general characteristics,

data source accessibility, data mining functionality and usability". A few other data mining tools like Rapid Miner, Weka, Orange and R were also reviewed in detail by Auza; (2010).

Rapid Miner which was previously popular as Yale can perform "process control, connect to a repository, import and export data, data transformation, modelling and evaluation". However, the open source version seems to only support CSV and MS Excel and have nil access to other databases (The Rapid Miner Platform; 2017)[3]. WEKA (Waikato Environment for Knowledge Analysis) on the other hand, is an open source software that extends its support to a gamut of visualization techniques and algorithms for analyzing data and predictive modelling, along with graphical user interfaces to provide ready access to such functions (Ian H. Witten and Hall; 2011). 'R' is yet another open-source data mining language and environment at one's disposal mainly used for statistical analysis, analytics and graphical representation (The R Project for Statistical Computing; 2017)[4].

Table 2: Comparison among different data mining tools

Algorithm	R	WEKA	RAPID MINER
Neural Network	✓	✓	✓
Decision Tree	✓	✓	✓
SVM	✓	✓	✓
k-NN	✓	✓	✓
Association Rule Mining	✓	✓	✓
Feature Selection	✓	✓	Hardly
Time Series Analysis	✓	✓	Hardly
Big Data Processing	✓	✓	Hardly

As is evident from Table 2, R and WEKA produces better results than Rapid Miner in many cases. Although R outperforms WEKA in various facets including handling missing values, visualization, supporting different data structures and analysis, WEKA can perform better in feature selection implementation (PEHLIVANLI; 2011). Thus, this research uses R as the analytical technique except for feature optimisation which is executed in WEKA.

Methodologies for data mining evaluation are majorly dependent on business applications, wherein one particular technique may prove to be more efficient than the other in view of the assessed application. Accordingly, the performance metrices in this research were calibrated for detection of malicious transactions. According to the contributions from literature (Jan; 2018; Yee and Saravanan, 2018; Yuanning Liu and Wang; 2011), the classifier performances are measured on the basis of accuracy, sensitivity and specificity. Hence, this study applies the same for assessing each of the actualised models.

3 Research Methodology

The methodology begins with determining the scope and purpose of the study in the relevant domain followed by data extraction and exploratory data analysis. Further, it focuses in preparing the data for experimental purpose, thereby interpreting the results of various hybrid algorithms to reach at the final discussion.

[3] The Rapid Miner Platform (2017). https://rapidminer.com/products/

[4] The R Project for Statistical Computing (2017). https://www.r-project.org/

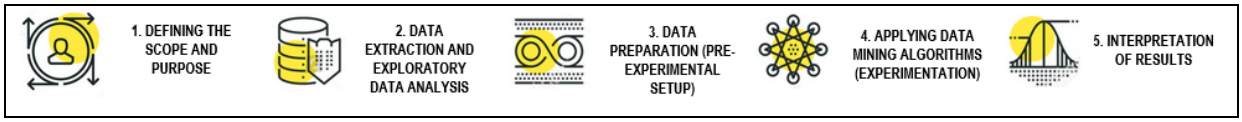

Figure 5: Research Methodology

3.1 Defining Scope and Purpose

The primary motive of this part is to comprehend the objectives and the need from the perspective of a business and then to convert this know-how into defining the problems of data mining. The business objective of the current research is to improve the detection of anomalous insurance claims in comparison to the existing in-vogue models in an attempt to reduce the considerable amount of losses that insurance companies are incurring globally.

3.2 Data Extraction and Exploratory Data Analysis

Here, an insurance fraud dataset is obtained from Kaggle to decipher the results of the hybrid model of feature optimisation and classification. Thereafter an exploratory data analysis is executed to gain a fair knowledge about the data such as understanding hidden or incomprehensible patterns or connections among features, correlation, detecting outliers and missing values, and most importantly identifying class imbalance (Martinez; 2010).

3.3 Data Preparation

This is the third and the most significant step of the data mining process for extracting the most effective outcome from the models to be experimented. Missing data is replaced with mode imputation for the categorical attributes (Sivagowry and Durairaj; 2014) and normalisation was performed to rescale the numeric variables within the range of 0 to 1 (Patro and Sahu; 2015). The data is then split into 70%-30% ratio for experimenting the hybrid algorithm – the greater part being used for training purpose and the smaller part for validation purpose.

3.4 Applying Data Mining Algorithms

This phase tests different hybrid models on the cleaned data. This study experiments with 10 different type of hybrid models of feature optimisation and classification. The feature optimisation techniques used here are Particle Swarm Optimisation (PSO) and Firefly Algorithm (FA) integrated with five different classifiers viz. Artificial Neural Network (ANN), Support Vector Machine (SVM), k- Nearest Neighbourhood (kNN), RandomForest (RF) and Naïve-Bayes (NB).

3.5 Interpretation of Results

In this section, the predictive ability of all the hybrid models is evaluated on the basis of accuracy, sensitivity and specificity (Yee; 2018), wherein sensitivity determines a fraudulent claim and specificity identifies fair transactions. Hence, this method helps to construct the architectural blueprint of an anomalous claim detection system to carry out research and development in the insurance sector.

4 Design Specification

Table 3 depicts the detailed pseudo code of the algorithm that has been proposed for the research and it aims to produce a comprehensive overview of the overall methodology.

Table 3: Pseudo code for the hybrid model of feature optimisation and classification

STEP		MODEL PSEUDO CODE
		Start Feature Selection
P	1	Generate initial particles and define constraints; set iteration = 0 and perform the subset selection process from step 2–4
S	2	Set iteration counter = iteration counter + 1.
	3	Calculate the fitness function for each particle and assign the best global position.
O	4	If stopping criteria is satisfied as per step 1, return the solution, else repeat step 2&3
		End Feature Selection
		Start Feature Selection
F	1	Generate initial population of fireflies and define constraints, set iteration = 1 and perform the subset selection process from step 2–4
A	2	Set iteration counter = iteration counter + 1.
	3	Calculate the fitness value of each firefly and assign the light intensity based on the objective function; Calculate the best position of the fireflies
	4	If stopping criteria is satisfied, return the solution, else repeat step 2-4.
		End Feature Selection
	5	Train the classifier (ANN, SVM, kNN, RF and NB) by the features obtained by PSO/Firefly.
	6	Measure accuracy, sensitivity and specificity for all the models.
	7	Analyse results and compare predictive ability of all the models generated.
		End Classification

5 Implementation

Here, the research implementation is illustrated. This quantitative study follows Cross Industry Standard Process for Data Mining (CRISP-DM) methodology against SEMMA and KDD (Daniel; 2005). The stages of CRISP-DM is the main motivation behind its usage as they could be duly constructed, arranged and characterised to enable effective comprehension and upgradation of any task.

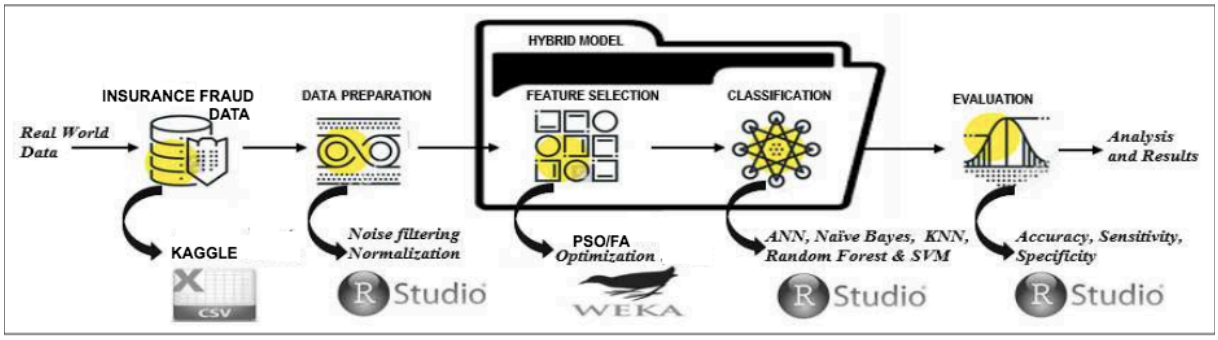

Figure 6: Research Methodology

5.1 Data Extraction and Exploratory Data Analysis

This research experiments the outcome of hybrid feature optimisation and classification on an insurance fraud dataset – acquired from Kaggle using R language. The data contains 39 features and 10211 instances out of which 5093 cases are of default (fraud_reported), which is found to be 49.8% of the total transactions. Missing values exist in three categorical variables, viz. 'collision type', 'property damage' and 'police report available' in the form of '?' which needs to be imputed for further analysis.

Figure 7: Extraction of Kaggle Insurance Claim Dataset

After extracting the data, exploratory data analysis (EDA) is performed to gather further understanding of the data, i.e class balance, correlation among variables, missing values and outliers. Correlation can be found between 'age' and 'months_as_a_customer' and also among other attributes viz 'total_claim_amount', 'injury_claim', 'property_claim,' and 'vehicle_claim'. A huge number of outlier is also evident from figure 8 which needs to be treated for better predictive performance (Sivagowry and Durairaj; 2014) of the models.

Figure 8: Boxplot showing outliers in the numeric variables

5.2 Data Preparation

The insurance dataset used for this study is comprised of 21 categorical and 18 numeric variables. There are missing values in the data for the categorical variables, and they are imputed with mode imputation with a view to attain an optimal model outcome (Sivagowry and Durairaj; 2014). The character variables are transformed into factors as and when required and the outliers are removed with the help of IQR (Inter quartile range) technique. For the next step, the data is normalised for re-scaling the numerical variables between 0 and 1. As a result, each input will have the same range of values (Patro and Sahu; 2015). The data is then split into 70%-30% ratio for experimenting the hybrid algorithm – the greater part being used for training purpose and the smaller part for validation purpose.

Figure 9: Splitting of data into train and test

5.3 Applying Data Mining Algorithms

The prepared data is experimented in three stages. All five classifiers have been individually combined with PSO and Firefly feature optimisation to create the hybrid model and the result is then compared with the one obtained by using Top 12 features by 'Random Forest Important Variable Selection Method'.
Further the feature optimisation is broken into three phases and the candidate-set acquired from this is used for filtering the data to train the model:

i Generating candidate-set that contains a subset of original attributes

ii Evaluating the candidate-set and estimating its utility and,

iii Determining the predictive potential of the specified variables.

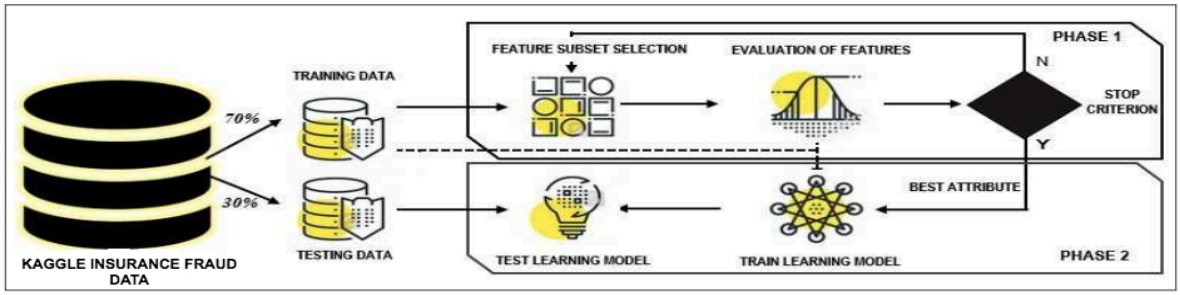

Figure 10: Consolidated view of the hybrid model

Particle Swarm Optimisation (PSO) and Firefly (FFA) feature optimisation technique is performed with the help of WEKA tool. PSO has curtailed the number of variables from 39 to 5 ('age', 'incident_state', 'incident_location', 'police_report_available' and 'injury_claim'), whereas, Firefly has considered 12 features ('policy_bind_date', 'insured_zip', 'policy_state', 'policy_annual_premium', 'incident_state', 'umbrella_limit', 'collision_type', 'insured_sex', 'insured_occupation', 'incident_location', 'property_damage' and 'bodily_injuried') as the most important ones among all.

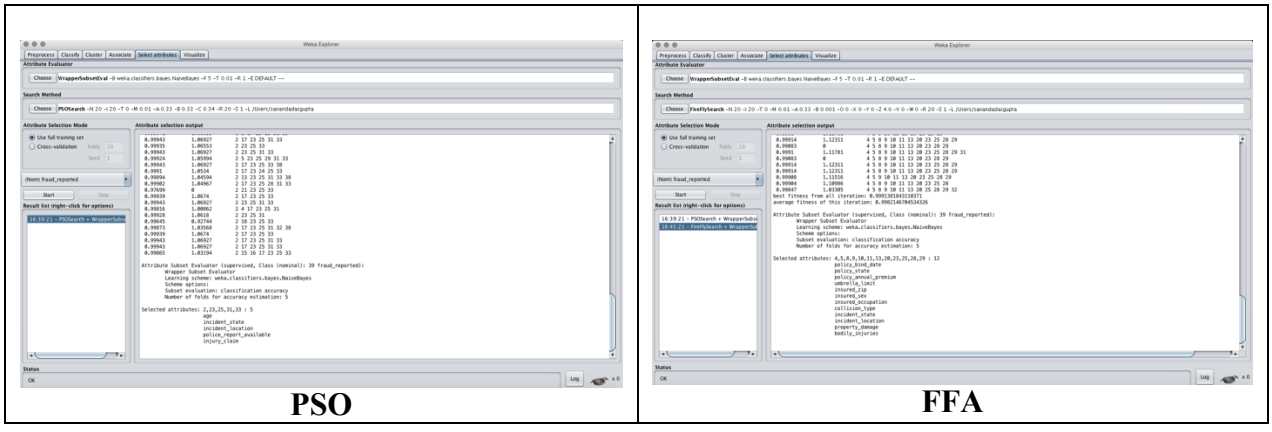

Figure 11: Implementation of Feature Optimisation in WEKA

The performance of the hybrid models is compared with the classifier performance when executed using top 12 important features ('insured_hobbies', 'incident_severity', 'incident_city', 'auto_make', 'auto_model', 'insured_occupation', 'insured_education_level', 'insured_relationship', 'incident_state', 'auhorities_contacted', 'policy_annual_premium') according to 'Random Forest Important Variable Selection' method.

```
#Top 12 feature selection using Random Forest Important Variable Selection Method
rf.model <- randomForest(fraud_reported ~ .,data=insurance_claim_updated, importance=TRUE,ntree=500)
rf.model
rf.1.var_imp <- varImpPlot(rf.model)
```

Figure 12: Top 12 features by Random Forest Important Variable Selection

5 and 12 optimal features obtained by PSO and FFA respectively are considered separately for the training of the classifiers in R- Studio. For example, Figure 13 depicts the experimentation of PSO combined with SVM classifier. It shows the usage of 'age', 'incident_state', 'incident_location', 'police_report_available' and 'injury_claim' for training the SVM classifier among all other 39 attributes.

```
#SVM Combined with PSO

svm_model=svm(formula = fraud_reported ~ .,
              data = trainTask,
              type = 'C-classification',
              kernel = 'linear')
svm_model
# Predicting the Test set results
svm_pred = predict(svm_model, newdata = testTask)
svm_pred
# Making the Confusion Matrix
cm = table(testTask$fraud_reported, svm_pred)
cm
result_svm <- confusionMatrix(testTask$fraud_reported, svm_pred)

```

Figure 13: Implementation of PSO combined with SVM

The remaining experiments (i.e. PSO_ANN, PSO_RF, PSO_NB, PSO_kNN, FFA_ANN, FFA_SVM, FFA_RF, FFA_NB and FFA_kNN) are conducted in a similar way to analyse the effectiveness of the model based on accuracy, sensitivity and specificity.

6 Evaluation

This section focuses on analysing the performance of 10 hybrid models viz. PSO combined with ANN, SVM, RF, NB, kNN and Firefly combined with the same five classifiers based on accuracy, sensitivity and specificity, as discussed in the literature review section. The output of these three performance metrices is calculated as illustrated in figure 14.

Predicted Class	**True/Actual Outcome**	
	Not Fraud	**Fraud**
Not Fraud / Fair Claims	**True Positive (TP)**	**False Positive (FP) Fair claims detected as fraud**
Fraud Claims	**False Negative (FN) Fraudulent claims identified as fair**	**True Negative (TN)**

$$Accuracy = \frac{(TP+FN)}{(TP+FP+FN+TN)} * 100$$

$$Sensitivity = \frac{(TP)}{(TP+FN)} * 100$$

$$Specificity = \frac{(TN)}{(TN+FP)} * 100$$

Figure 14: Confusion Matrix and Formula of Performance Metrices for Binary Classification

6.1 Experiment 1 (PSO_SVM)

The first experiment is executed to summarise the performance of hybrid SVM and PSO. According to Table 4, the model correctly classifies 1664 instances with accuracy of 54%, however 1400 observations are incorrectly identified by the model. It attains a sensitivity of 53% by perfectly detecting 931 suspicious claims out of 1766 transactions, whereas, records specificity of 56% by incorrectly noting 565 transactions as fraud out of 1298 claims.

Table 4: Confusion Matrix and Performance Evaluation of PSO_SVM

		Actual	
Predicted		Fraud	Not Fraud
	Fraud	931	565
	Not Fraud	835	733

Performance Metric	Value
Accuracy	0.54
Sensitivity	0.53
Specificity	0.56

6.2 Experiment 2 (PSO_RF)

The second experiment demonstrates the outcome of hybrid RF and PSO. In this case the model accurately classifies 1766 instances with 57% accuracy, whereas, it incorrectly classifies 1298 transactions as depicted by Table 5. The model provides sensitivity of 56% by correctly identifying 897 cases of fraud out of 1596 observations. However, it mistakenly shows 699 transactions as fraudulent which are not true. On the other hand, the model correctly detects 869 claims as fair out of 1468 claims and 599 claims are identified incorrectly as fraud, recording the specificity as 59%.

Table 5: Confusion Matrix and Performance Evaluation of PSO_RF

		Actual	
Predicted		Fraud	Not Fraud
	Fraud	897	599
	Not Fraud	699	869

Performance Metric	Value
Accuracy	0.57
Sensitivity	0.56
Specificity	0.59

6.3 Experiment 3 (PSO_NB)

The third experiment is carried out to comprehend the result of hybrid NB and PSO. According to Table 6, the model correctly classifies 1764 instances with accuracy of 58%, however the model provides 1300 incorrect findings. It attains a sensitivity of 56% by accurately indicating 925 suspicious claims out of 1654 transactions, whereas, yields a specificity of 59% by incorrectly recording 571 transactions as fraud out of 1410 claims.

Table 6: Confusion Matrix and Performance Evaluation of PSO_NB

		Actual	
		Fraud	**Not Fraud**
Predicted	**Fraud**	925	571
	Not Fraud	729	839

Performance Metric	**Value**
Accuracy	0.58
Sensitivity	0.56
Specificity	0.59

6.4 Experiment 4 (PSO_kNN)

The fourth experiment determines the performance of hybrid kNN and PSO. Here, the model reliably classifies 1877 instances with 61% accuracy, whereas, it incorrectly classifies 1187 transactions as outlined in Table 7. The model gives sensitivity of 60% by accurately identifying 912 cases of fraud out of 1496 observations. However, it improperly records 584 transactions as fraudulent which are not true. Besides, the model correctly detects 965 claims as fair out of 1568 claims and 603 claims are incorrectly identified as fraud, recording a specificity of 62%.

Table 7: Confusion Matrix and Performance Evaluation of PSO_kNN

		Actual	
		Fraud	**Not Fraud**
Predicted	**Fraud**	912	603
	Not Fraud	584	965

Performance Metric	**Value**
Accuracy	0.61
Sensitivity	0.60
Specificity	0.62

6.5 Experiment 5 (PSO_ANN)

The fifth experiment is executed to interpret the result of hybrid ANN and PSO. According to Table 8, the model correctly classifies all 3064 instances with an accuracy of 100%. It attains

a sensitivity of 1 by accurately detecting all 1496 suspicious claims, along with a specificity of 1 by immaculately recording all 1568 fair transactions.

Table 8: Confusion Matrix and Performance Evaluation of PSO_ANN

		Actual	
		Fraud	**Not Fraud**
Predicted	**Fraud**	1496	0
	Not Fraud	0	1568

Performance Metric	Value
Accuracy	1.00
Sensitivity	1.00
Specificity	1.00

6.6 Experiment 6 (FFA_SVM)

The sixth experiment depicts the performance of hybrid FFA and SVM. Here, the model reliably classifies 1761 instances with 61% accuracy, whereas, it incorrectly classifies 1303 transactions as depicted by Table 9. The model gives sensitivity of 60% by accurately identifying 823 cases of fraud out of 1453 observations. However, it mistakenly shows 630 transactions as fair which are not true. Besides, the model correctly detects 938 claims as fair out of 1611 claims and 673 claims are identified incorrectly as fraud, recording the specificity as 62%.

Table 9: Confusion Matrix and Performance Evaluation of FFA_SVM

		Actual	
		Fraud	**Not Fraud**
Predicted	**Fraud**	823	673
	Not Fraud	630	938

Performance Metric	Value
Accuracy	0.57
Sensitivity	0.57
Specificity	0.58

6.7 Experiment 7 (FFA_RF)

The seventh experiment demonstrates the result of hybrid RF and FFA. According to Table 10, the model correctly classifies 2974 instances with accuracy of 97%, however 90 observations are incorrectly identified by the model. It attains a sensitivity of 97% by accurately detecting 1445 suspicious claims out of 1496 transactions, whereas, it records a specificity of 96% by incorrectly recording 39 transactions as fraud out of 1568 claims.

Table 10: Confusion Matrix and Performance Evaluation of FFA_RF

		Actual	
Predicted		Fraud	Not Fraud
	Fraud	1445	39
	Not Fraud	51	1529

Performance Metric	Value
Accuracy	0.97
Sensitivity	0.97
Specificity	0.96

6.8 Experiment 8 (FFA_NB)

The eighth experiment illustrates the performance of hybrid FFA and NB. Here, the model properly classifies 1758 instances with 57% accuracy, whereas, it incorrectly classifies 1306 transactions as depicted by Table 11. The model gives sensitivity of 56% by accurately identifying 892 cases of fraud out of 1496 observations. However, it mistakenly shows 604 transactions as fair which are not true. Besides, the model correctly detects 866 claims as fair out of 1568 claims and 702 claims are identified incorrectly as fraud, recording a specificity of 59%.

Table 11: Confusion Matrix and Performance Evaluation of FFA_NB

		Actual	
Predicted		Fraud	Not Fraud
	Fraud	892	702
	Not Fraud	604	866

Performance Metric	Value
Accuracy	0.57
Sensitivity	0.56
Specificity	0.59

6.9 Experiment 9 (FFA_kNN)

The ninth experiment exhibits the result of hybrid RF and FFA. According to Table 12, the model correctly classifies 2934 instances with accuracy of 96%, however 130 observations are incorrectly identified by the model. It attains a sensitivity of 99% by accurately detecting 1376 suspicious claims out of 1496 transactions, whereas, it records specificity of 93% by incorrectly recording 10 transactions as fraud out of 1568 claims.

Table 12: Confusion Matrix and Performance Evaluation of FFA_kNN

		Actual	
Predicted		Fraud	Not Fraud
	Fraud	1376	10
	Not Fraud	120	1558

Performance Metric	Value
Accuracy	0.96
Sensitivity	0.99
Specificity	0.93

6.10 Experiment 10 (FFA_ANN)

The tenth experiment is carried out to comprehend the result of hybrid ANN and PSO. According to Table 13, the model correctly classifies all 3064 instances with accuracy of 100%. It attains a sensitivity of 1 by accurately detecting all 1496 suspicious transactions, along with a specificity of 1 by immaculately recording all 1568 fair transactions.

Table 13: Confusion Matrix and Performance Evaluation of FFA_ANN

		Actual	
Predicted		Fraud	Not Fraud
	Fraud	1496	0
	Not Fraud	0	1568

Performance Metric	Value
Accuracy	1.00
Sensitivity	1.00
Specificity	1.00

6.11 Discussion

An elaborate review of the outcomes from the 10 experiments are illustrated in this section. For better understanding of the results obtained from the study, the research is also conducted using a few techniques (viz. SVM, RF and NB) as mentioned in Table 1 by Seeja *et al.* (2014). Apart from PSO and FFA, 12 important variables are identified and applied on all five classifiers to check the change in their performances and it is worth mentioning that the output has outperformed the hybrid of PSO combined with all 5 classifiers. When compared to the current state of art in Table 1, the results obtained in this study for SVM, RF and NB have topped the list with higher accuracy, sensitivity and specificity (Table 14).

Table 14: Performance Evaluation of SVM, RF and NB as compared to Table 1

Performance Metric			
Model	**Accuracy**	**Sensitvity**	**Specificity**
SVM (Seeja *et al*., 2014)	0.55	0.60	0.55
SVM (Current Study)	0.86	0.87	0.85
RF (Seeja *et al.*, 2014)	0.80	0.78	0.72
RF (Current Study)	0.96	0.92	0.99
NB (Seeja *et al.*, 2014)	0.60	0.63	0.54
NB (Current Study)	0.80	0.82	0.78

In this research, accuracy is the measure of correct fraud predictions, whereas, sensitivity shows how well the model can identify the actual fraud cases against total fraud claims. Specificity is the true negative rate which determines how well the model can identify the fair claims.

As evident from Table 15, the results of hybrid PSO and all 4 classifiers, except ANN, are quite disappointing in terms of all three performance metrices. ANN performed equally well for both the feature optimisation techniques. The output obtained from ANN are quite unrealistic and can be a result of overfitting of the model, which is quite common for neural networks. The overfitting might have occurred due to the existence of noise in the data, hence the performance of the validation set is much lower than the performance of the training data. As the model could not generalise well the error on training set is much lower than the test set. Therefore, the paper can't conclude ANN as the best performing classifier among all, rather requires an in-depth future analysis to identify and prevent the problem of overfitting. On the other hand, FFA combined with RF is leading other 3 classifiers followed by FFA_kNN in terms of accuracy and specificity. Though FFA_kNN (0.99) has achieved higher sensitivity than FFA_RF (0.97), the false negative rate of the prior (0.08%) is higher than the later (0.03%). Though kNN can produce higher accuracy but it is not competitive in comparison to RF, as it learns nothing from the training data but only uses this for the purpose of classification.

Table 15: Performance Comparison of 10 hybrid models

Performance Metric			
Model	**Accuracy**	**Sensitvity**	**Specificity**
PSO_SVM	0.54	0.53	0.56
FFA_SVM	0.57	0.57	0.58
PSO_RF	0.57	0.56	0.59
FFA_RF	0.97	0.97	0.96
PSO_NB	0.58	0.56	0.59
FFA_NB	0.57	0.56	0.59
PSO_kNN	0.61	0.60	0.62
FFA_kNN	0.96	0.99	0.93
PSO_ANN	1	1	1
FFA_ANN	1	1	1

Therefrom, the research concludes that a hybrid combination of Firefly Algorithm (FFA) and Random Forest (RF) is the most effective algorithm in detecting anomalous claims within the

insurance domain and can definitely bring significant improvement in predictive performance of the model when compared to the other three hybrid algorithms applied in this study.

7 Conclusion and Future Work

The purpose of this study is to check the predictive capability of a hybrid of feature optimisation and classification in discerning suspicious transactions in the insurance industry and whether it can bring any effective advancement over the current state of art. Two optimisation techniques (PSO and FFA) along with five classification algorithms (ANN, SVM, RF, NB and kNN) are chosen to conduct the research and it is visible from the output that the results of FFA combined with RF has outperformed the current state of art when experimented on the insurance claim dataset extracted from Kaggle. The evaluation of the learning model is based on accuracy, sensitivity and specificity. According to the current scenario, deep learning methods like ANN has struggled due to the existence of numerous high cardinality categorical variables in the data as it ends up in building infinitely wide neural net making the process extremely slow. The study also outperformed some techniques mentioned in the literature review section (SVM, RF and NB) and this may be due to usage of different data preparation and feature selection techniques. Another drawback of this research is that hybrid model has hard to no practical implementation or there is no real time scenario that will predict into future. It can only delve into learning what has already happened.

Looking at the task from a classification perspective a realistic result may possibly be achieved from ANN by applying binary or hashing encoders, which may be appropriate to encode all of the categories into a single representation per feature vector and not allowing any single one to dominate the other. Embedding can also be used to transform large number of categorical variables into a single vector. PCA-CAT can be another worth trying feature engineering technique in the future. To make this process faster, random search is chosen for hyper parameter tuning procedure. However, it may overlook some important combination of parameters which can produce more accuracy. It is for the reason that grid search can be used for the future as a different hyper parameter tuning technique as it considers all possible combination of the parameters. Few other metaheuristic approaches like 'Simulated Annealing', 'Ant Colony Optimisation' can be attempted in the future for optimising various features which may generate better performance from the models. A special type of polythetic decision tree – ART as well as few other machine learning algorithms like GBM, XGboost can also be worth trying hereafter. As future work, performance of all the hybrid algorithms can be tested on few other as well as on real time datasets to verify the predictive capability of the models.

References

Adaniya, M., Abr˜ao, T. and Proenc¸a Jr., M. (2013). Anomaly Detection Using Metaheuristic Firefly Harmonic Clustering. Journal of Networks, 8(1): 1183-1187.

Arora, S. and Kumar, D. (2017). Hybridization of SOM and PSO for Detecting Fraud in Credit Card. International Journal of Information Systems in the Service Sector, 9(3): 17-36.

Auza, J. (2010). 5 of the best and free open source data mining software.
URL: http://www.junauza.com/2010/11/free-data-mining-software.html

Barrera, R., Gómez, I. and Quiroga, J. (2014). Structural damage detection: comparison between GA and PSO techniques. 29(1): 61-70.

Benchaji I., Douzi S., El Ouahidi B. (2019) Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection, 66, Springer.

Bodaghi, A., & Teimourpour, B. (2018). The detection of professional fraud in automobile insurance using social network analysis.

Bramer, M. (2013). Principles of Data Mining. 2nd edition, Springer.

B.N.Lakshmi and G.H.Raghunandhan (2011). A conceptual overview of data mining: 27–32.

Camarda, L., D'Arienzo, A., Grassedonio, E., Zerbo, S., Argo, A. and D'Arienzo, M. (2018). Self-inflicted long bone fractures for insurance fraud. International Journal of Legal Medicine.

Daniel, L. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley Sons, Inc., Hoboken, New Jersey.

D. P. Shukla, S. B. P. and Sen, A. K. (2014). A literature review in health informatics using data mining techniques, International Journal of Software Hardware Research in Engineering.

Erdinc, O. (2017). Optimization in Renewable Energy Systems. Elsevier.

Ghorbani, R & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some disease's prediction. International Journal of Data and Network Science, 3(2): 47-70.

Herland, M., Baude, R. and Khoshgoftaar, T. (2017). Medical Provider Specialty Predictions for the Detection of Anomalous Medicare Insurance Claims. 2017 IEEE International Conference on Information Reuse and Integration (IRI), 579-588.

Huan Liu, Hiroshi Motoda, R. S. and Zhao, Z. (2010). Feature selection: An everlasting frontier in data mining, JMLR: The 4th Workshop on Feature Selection and Data Mining.

Ian H. Witten, E. F. and Hall, M. A. (2011). Data Mining: Practical machine learning tools and techniques, 3rd Edition, Morgan Kaufmann, San Francisco.

Jan, C.L. (2018). An Effective Financial Statements Fraud Detection Model for the Sustainable Development of Financial Markets: Evidence from Taiwan, 10: 513-537.

Jamjala, S., Thakur, P., Rohra, J.G., Bhatt, R.B. and Perumal, B. (2018). User Localisation in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithmwith Neural Networks, 2: 286-295.

Kirildog, M. and Asuk, C. (2012). A Fraud Detection Approach with Data Mining in Health Insurance. Procedia - Social and Behavioral Sciences, 62: 989-994.

Kose, I., Goktur, M. and Kilic, K. (2015). An interactive machine- learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing, 36: 283-299.

Larnyo, E., Udimal, T. (2018). Detecting and Combating Fraudulent Health Insurance Claims Using ANN. Journal of Health, Medicine and Nursing.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R., Tang, J. and Liu, H. (2017). Feature Selection-A Data Perspective. ACM Computing Surveys, 50(6): 1- 45.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification, IEEE.

Marjia Sultana, A. H. and ShorifUddin, M. (2016). Analysis of data mining techniques for heart disease prediction, IEEE.

Martinez, W. L.; Martinez, A. R. . S. J. (2010). Exploratory Data Analysis with MATLAB, second edition, Chapman Hall/CRC.

Mohammed Abdul Khaled, S. K. P. and Dash, G. (2013). A survey of data mining techniques on medical data for finding locally frequent diseases, International Journal of Advanced Research in Computer Science and Software Engineering, 1137–1145.

Mondal, S. (2017) Diagnosis of Cardiovascular Diseases using Hybrid Feature Selection and Classification Algorithms. Master's thesis, Dublin, National College of Ireland.

Ngai, E., Hu, Y., Wong, Y., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3): 559-569.

Nian, K., Zhang, H., Tayal, A., Coleman, T. and Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1): 58-75.

Patro, S. G. K. and Sahu, K. K. (2015). A technical analysis of financial forecasting, International Journal of Computer Sciences and Engineering.

PEHLİVANLI, D. A. (2011). The comparison of data mining tools, Department of Computer Engineering, İstanbul Kültür University.

Pilbeam, K., (2018). Finance & financial markets. London: Macmillan International Higher Education.

Priya, K. U., & Pushpa, S. (2017). A Survey on Fraud analytics using Predictive Model in Insurance Claims. International Journal of Pure and Applied Mathematics, 114(7): 755-767.

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE access, 6, 14277-14284.

Rish, I. (2001). An empirical study of the naive bayes classifier, IJCAI Workshop on Empirical Methods in AI.

Sivagowry, S. and Durairaj, M. (2014). An intellectual technique for feature reduction on heart malady anti- cipation data, International Journal of Advanced Research in Computer Science and Software Engineering.

Sönmez, F., Zontul, M., Kaynar, O., & Tutar, H. (2018). Anomaly Detection Using Data Mining Methods in IT Systems: A Decision Support Application. Sakarya University Journal of Science, 22(4): 1109-1123.

Subudhi, S. and Panigrahi, S. (2016). Use of fuzzy clustering and support vector machine for detecting fraud in mobile telecommunication networks. International Journal of Security and Networks, 11(1/2): 3.

Sundarkumar, G. and Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence, 37: 368-377.

Shwartz, S. and David, B. (2014). Understanding Machine Learning, Cambridge University Press, New York.

Swathi Jamjala Narayanan, Boominathan Perumal and Jayant G., (2018). Modeling, Analysis, and Application of Nature-Inspired Metaheuristic Algorithms.

Tayal, K., Ravi, V. (2016). Particle swarm optimization trained class association rule mining: Application to phishing detection. Proceedings of the International Conference on Informatics and Analytics, 6: 13-23.

Tyagi, H. and Rakesh, N. (2018). Enhanced Online Hybrid Model for Online Fraud Prevention and Detection. Proceedings of First International Conference on Smart System, Innovations and Computing, 97-106.

Umarani, R., Sithic, H.L. (2015). Fuzzy Matrix Theory as a Knowledge Discovery in Health Care Domain. Procedia Computer Science, 47: 282-291.

Viaene, S., Derrig, R., Baesens, B. and Dedene, G. (2002). A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection.Journal of Risk & Insurance, 69(3): 373-421.

Verma, A., Taneja, A. and Arora, A. (2017). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. 2017 Tenth International Conference on Contemporary Computing (IC3).

Wahid A., Rao A.C.S. (2019) A Distance-Based Outlier Detection Using Particle Swarm Optimization Technique, 40, Springer.

Warren, D. and Schweitzer, M. (2018). When Lying Does Not Pay: How Experts Detect Insurance Fraud. Journal of Business Ethics, 150(3): 711-726.

Webster, J. and Watson, R. T. (2002). Analyzing the past to preparefor the future:writing aliterature review, MIS, 1137–1145.

West, J., and Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57: 47-66.

X. Chen, Y. Y. and Williams, G. (2007). A survey of open source data mining systems emerging technologies in knowledge discovery and data mining, Springer.

Yao, J., Zhang, J. and Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, 57-61.

Yee, O.S., Saravanan, S. (2018). Credit Card Fraud Detection Using Machine Learning As Data Mining Technique. Elsevier, 32: 392-423.

Yuanning Liu, G. W. and Wang, S. (2011). An improved particle swarm optimization for feature selection, Journal of Bionic Engineering.

Zhang, R., Nie, F., Li, X. and Wei, X. (2019). Feature selection with multi-view data: A survey. Information Fusion, 50: 158-167.

Zareapoor, M. and KR, S. (2014). Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection. International Journal of Information Engineering and Electronic Business, 7: 60-65.