International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 257-265

Modeling Insurance Fraud Detection Using Ensemble Combining Classification

Amira Kamil Ibrahim Hassan[1] and Ajith Abraham[2]

[1] Management Information Systems Department, School of Management, Ahfad University for Women
Department of computer science, Sudan University of Science and Technology,
Khartoum, Sudan
amirakamil2@yahoo.com

[2] Machine Intelligence Research Labs (MIR Labs), WA, USA
IT4Innovations, VSB - Technical University of Ostrava, Czech Republic
ajith.abraham@ieee.org

***Abstract*: This paper is a continuation of previous paper where the imbalance dataset problem was solved by applying a proposed novel partitioning-undersampling technique. Then a proposed innovative Insurance Fraud Detection (IFD) models were designed using base-classifiers; Decision Tree, Support Vector Machine and Artificial Neural Network. This paper proposed an innovative insurance fraud detection models by applying ensemble combining classifiers on IFD models designed previously using base-classifiers. Throughout the paper, ten-fold cross validation method of testing is used. Its originality lies in the use of several ensembles combining classifier and comparing between them for choosing the best model. Results from a publicly available automobile insurance fraud detection dataset demonstrate that DTIFD performs slightly better than all proposed models, ensemble combining classifier designed IFD models with high recall but still DTIFD model was the best. The proposed models were applied on another imbalance datasets and compared. Empirical results illustrate that the proposed models gave better results.**

***Keywords*: Insurance fraud detection, imbalanced data, Voting, Stacking and Grading.**

I. Introduction

Insurance fraud is a significant and costly problem for both policyholders and insurance companies in all sectors of the insurance industry. In recent years, fraud detection has attracted a great deal of concern and attention. The Oxford English Dictionary [1] defines fraud as "wrongful or criminal deception intended to result in financial or personal gain".

Fraud occurs in a wide variety of forms and is ever changing as new technologies and new economic and social systems provide new opportunities for fraudulent activity. The total extent of business losses due to fraudulent activities is difficult to define. Phua et al. [2] described fraud as leading to the abuse of a profit organization's system without necessarily leading to direct legal consequences. Although there is no universally accepted definition of financial fraud, Wang et al. [3], defined it as "a deliberate act that is contrary to law, rule, or policy with intent to obtain unauthorized financial benefit". Economically, insurance fraud is becoming an increasingly serious problem.

Insurance fraud detection (IFD) is important for preventing the disturbing results of insurance fraud. IFD involves distinguishing fraudulent claims from genuine claims, thereby disclosing fraudulent behavior or activities and enabling decision makers to develop appropriate strategies to decrease the impact of fraud.

Fraud is a major problem causing a lot of losses for many insurance companies. Data mining can minimize some of these losses by making use of the massive collections of customer data. Besides scalability and effectiveness, the fraud-detection task is faced with technical problems that include imbalanced dataset, which have not been widely studied in the insurance fraud detection community. The insurance fraud detection or generally fraud detection data is imbalanced. The fraudulent cases are minority class while the legitimate cases are big majority class. Using the data as it is results in high success rate for predicting legitimate cases but without predicting any fraudulent cases. There are two methods that are used to solve this problem [4].

Since ensemble learning attracts much attention from pattern recognition and machine learning for good performance, several ensemble methods has applied to insurance fraud detection field in general but not that much in the automobile insurance fraud detection [5].

This paper introduces the new fraud detection method by using dataset that was re-sampled by using a proposed partitioning-undersampling technique. The decision tree, support vector machine and artificial neural network classifiers were used to design IFD models named base-classifiers models though out this research. The innovative use of grading, stacking and voting to process the IFD models has the possibility of getting better results. One related problem caused by imbalanced data includes measuring the performance of the classifiers. Recent work on imbalanced data sets was evaluated using better performance

metrics such as recall, precision [6] [7] [8] and area under curve [9]. In this paper, recall, precision and PRC Curves (the Precision-Recall Characteristic curve) were used to evaluate the performance of models. Section 2 contains literature review, mentioning all recent researches in the area of insurance fraud detection and also Section 3contains explains the data partitioning technique applied and includes also a brief description of data mining algorithms used. Section 4 contains data pre-processing and experimental setup and Section 5 illustrates the results and analysis followed by Conclusions.

II. Literature Review

A. Introduction:

The literature review is divided into two main topics. The first topic is Insurance fraud detection with especial interest in automobile insurance fraud detection. The second topic deals with techniques used for solving imbalance dataset problem.

B. Insurance Fraud Detection:

Data mining has a significant role in IFD, as it is often applied to extract and uncover the hidden truths behind very large quantities of data. Data mining is about finding insights which are statistically reliable, unknown previously, and actionable from data [10]. This data must be available, relevant, adequate, and clean. Also, the data mining problem must be well-defined, cannot be solved by query and reporting tools, and guided by a data mining process model [11].

Bose and Mahapatra [12] defined data mining as a process of identifying interesting patterns in databases that can then be used in decision making. Turban et al. [13] defined data mining as a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequently gain knowledge from a large database. Frawley et al. stated that the objective of data mining is to obtain useful, non-explicit information from data stored in large repositories [14]. Kou et al. highlighted that an important advantage of data mining is that it can be used to develop a new class of models to identify new attacks before they can be detected by human experts [15]. Phua et al. pointed out that fraud detection has become one of the best established applications of data mining in both industry and government [2]. Various data mining techniques have been applied in IFD, such as artificial neural networks, logistic regression models, Naïve Bayes, support vector machine method and decision trees, among others [2].

The data mining techniques used for insurance fraud detection in published academic papers were identified; those papers were classified according to the used data mining techniques. All published papers on insurance fraud detection (IFD) using data mining technique in the period between 1997 and 2015 were classified [16]. The search phrase used was "insurance fraud detection data mining", the search was done in the time period between 1997 and 2013 first, and then the same search was done for years 2013 to 2015. The detail results of this search was published in previous paper [17].

The data mining techniques used in insurance fraud detection are classified into six data mining application classes of classification, clustering, prediction, outlier detection, regression, and visualization. The insurance fraud consists of three types: automobile insurance (AI), crop insurance (CI) and healthcare insurance (HI). It was noticed that the most used data mining application class in all three insurance fraud types is classification thus it is used in this research.

Then more categorization is done using data mining algorithmic used (e.g., neural networks). Classification of the 44 papers according to data mining techniques, illustrated that the most often used techniques are logistic models, the Naïve Bayes, Decision tree, support vector machine, and artificial neural network all of which fall into the "classification" class. The majority of these papers are in automobile insurance fraud detection, it is believed that it has to do with the data collection. It very difficult to collect data for insurance fraud detection in general but there is punch-data available for automobile fraud detection.

In this research the base classifiers used are artificial neural network, decision tree and support vector machine. This choice was done based on the result of the review of the past papers. The most used classifier Logistic model and Naïve Bayes were not chosen because it would be a repeating for previous work. The classifiers were chosen from the ten most used to insure good results. Another criteria for choosing those classifiers are shown in details in previously published paper [18].

C. Techniques for Solving Imbalanced Dataset Problem:

When applying conventional machine learning algorithms the mining of imbalanced datasets can result in models that are strongly predictive for the larger class, while delivering performance which is poorly predictive for the minority class [19] [20]. This is due to the fact that conventional classifiers will attempt to return the most correct predictions based upon the entire dataset, this results in them categorizing all data as belonging to the larger class. This class is usually the class, which is of least interest to the data-mining problem. In the case of insurance fraud data mining, the majority class is the class where no fraud has occurred and the minority class being fraud. When minority class is very small a learner can deliver very high predictive accuracy even though it has classified none of the minority class correctly. Taking the area of interest of this thesis (fraud) the minority class would be 'possibly fraudulent claims' while the majority class would be 'non fraudulent'.

Several classification techniques have been proposed and applied in the literature for imbalanced classification problems. These techniques can be classified in two major categories: resampling-then-classification and cost-sensitive learning. However, there are ensemble algorithms, which build an ingratiation of classifiers. Typically, these algorithms are ensemble of cost-sensitive learning or resampling-then-classification algorithms. The objective of using ensemble learning is to improve the classification performance. In this research a novel proposed resampling technique is used then ensemble techniques to design improved IFD models.

Ensemble Learning

The main idea of ensemble methodology is to combine a set of models, each of which solves the same original task, in order to obtain a better composite global model, with more accurate and reliable estimates or decisions than can be obtained from using a single model [21]. There are three main types Sequential Methodology, Concurrent Methodology and

Combining Classifiers. The ensemble models designed in this research are designed using Combining Classifiers.

Combining Classifiers way of combining the classifiers may be divided into two main groups: simple multiple classifier combinations and meta-combiners. The simple combining methods are best suited for problems where the individual classifiers perform the same task and have comparable success. However, such combiners are more vulnerable to outliers and to unevenly performing classifiers. On the other hand, the meta-combiners are theoretically more powerful but are susceptible to all the problems associated with the added learning (such as over-fitting, long training time). In this research the simple combining method used is averaging and voting. The Meta combining methods used is stacking and grading.

A typical performance measure for classification is the so-called accuracy, which is calculated as the correctly classified samples over the total number of training samples. However, for imbalanced classification problems this might not be a good performance indicator, since the majority class dominates the behavior of this metric. More specifically, naive decision rules can yield high classification accuracy. Alternatively, recall and precision can be used. They are defined as

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TN+FP} \quad (2)$$

Still, precision is manipulated by the majority (negative) class. However, the recall is not and therefore, it is a more appropriate measure for this purpose. The space spanned by recall and precision is termed recall (x-axis)/Precision (y-axis) curve (PRC area). The PRC area provides a good visual representation. In this research, recall, precision and PRC are used as performance measures.

Special concentration is given to published papers in the area of automobile insurance fraud detection (AIFD), which used imbalanced dataset. The main problems facing IFD are imbalanced data and the choosing of data mining classifiers that give the best results. Analyzing AIFD past papers, mention in previous review paper [17] and few recent papers, reviewing the methods used to treat the imbalance data problem, DM classifier used and the evaluation methods used. The aim of analyzing the DM technique and imbalance dataset problem solving techniques is to show the unique of the proposed methods in this research.

In the following the techniques used in those papers, to solve the imbalance dataset problem are mentioned. Sternberg and Reynolds solved the problem by search manually for the features that cause type 1error (false positive) and type 2 error (false negative), and use these features to design the model [22]. Brockett et al., and Tennyson and Salsas Forn; improved the method a little by sorting the data into categories by an expert, this way the whole dataset is used [23], [24]. Caudill et al. and Artís et al. in two papers, used an oversampling of fraud claims in order to obtain a good representation for this group [25] [26] [27]. Pérez et al. also used oversampling of the fraud claims but testing with different percentage [6]. Other technique used to solve the imbalance data problem is random sampling with all its types. Belhadji et al. used simple random sampling [28]. Pinquet et al. divide the dataset randomly into two Holdout sample (Random auditing) and Working sample (Usual auditing strategy) [29]. Derrig et al. and Farquad et al. randomly resample the dataset using a stratified sampling using blocked ten-fold cross-validation [30] [7].

Another method used in few papers is partitioning the dataset into several subsamples Viaene et al. partitioning-sample, repeated 100 times, each time using a different randomization selection of the data [31]. Then Viaene et al. improved the above method by resampling the dataset by randomly partitioning the data into k disjoint sets of approximately equal size, and then use k fold cross validation [32]. Xu et al. generated multiple training subsets in terms of the reductions produced by rough set reduction technique [33]. Vasu and Ravi, proposed a hybrid undersampling approach that employs k-reverse nearest neighbour (kRNN) method to detect the outliers from majority class then using K-means clustering to further reduce the influence of the majority class [34]. Sundarkumar and Ravi further improved the proposed sampling method by including One-class support vector machine to reduce the majority class even more [8].

Those papers mentioned above were reviewed for the data mining algorithms used.

Sternberg and Reynolds, designed fraud detection expert system using the Cultural Algorithms (CA) that provides self-adaptive capabilities, which can generate the information necessary for the expert system to respond dynamically and provide an automated response to environmental changes [22]. Brockett et al. apply Kolionen's used self-organizing feature map to classify automobile bodily injury claims by degree of fraud suspicion, neural network and a back propagation were used to investigate the validity of feature map approach and showed that it performed better than previous methods [23]. Several researches used logit model [25], [26], [24], [30], [27], [35]. Neural network was used by [30] [32]. Xu et al. (2011) used Neural network classifier then improved the performance of the classifier by designing an ensemble neural network [33]. Support vector machine is used in several papers [7] [30] [8] [34] and decision tree is also used in a lot of papers [30] [6] [36] [34] [7] [8]. Belhadji et al. and Pinquet et al. used Probit model to design expert system, then improved it more by using a Probit model, a two equation model for audit and fraud (a Bivariate Probit model with censoring) was estimated on a sample of suspicious claims for which the experts were left to take the audit decision. Results were rather close to those obtained with a random auditing strategy, at the expense of some instability with respect to the regression components set [28] [29] .

Brockett et al. used another statistical technique principal component analysis of RIDIT (PRIDIT) [37]. Pathak et al. (2005) developed a fuzzy logic based expert system that can identify and evaluate whether elements of fraud are involved in insurance claims settlement thus reduce the need for human experts [38].

Viaene et al. compared between several data mining classifiers, Neural networks, support vector machine (SVM), K-nearest nieghbor, Naïve Bayes (NB), Bayesian belief network, decision trees and Logistic model (LM), the results show that there is no great difference between the different classifiers, except SVM, LM and NB result were a little better [30]. Viaene et al. used (smoothed) naive Bayes (NB), AdaBoosted naive Bayes (AB), and AdaBoosted weights of evidence (ABWOE) comparing them on this real-life data set, the boosted weight of evidence algorithm showed comparable (slightly better) discriminatory and ranking ability to (smoothed) naive Bayes with and without boosting, but clearly improved on the calibration of probability estimates [31]. Pérez et al. compared between two decision tree induction algorithms C4.5 and CTC [6] . Bhowmik et al., compared between Naïve Bayesian classification and decision tree-based classification using two decision tree algorithms (C4.5 algorithms and Consolidated Trees) [36]. Farquad et al. used SVM-RFE for feature selection, and for rule generation used decision tree (DT) and Naive Bayes tree (NB Tree) [7]. Vasu and Ravi compared between several classifiers, support vector machine (SVM), logistic regression (LR), multi layer perceptron (MLP), radial basis function network (RBF), group method of data handling (GMDH), genetic programming (GP) and decision tree (J48) [34]. Sundarkumar and Ravi, improved the previous propped undersampling method tested its efficiency by comparing the new results with previous result, again using several classifier algorithms support vector machine (SVM), logistic regression (LR), multi layer perceptron (MLP), radial basis function network (RBF), group method of data handling (GMDH), genetic programming (GP) and decision tree (J48); it was shown that the new proposed method gave better results; DT and SVM produced the best results [8].

III. Background

A. Data Partitioning

According to Chan, et al. [39], the desired distribution of the data partitions belonging to a particular fraud detection data set must be determined experimentally. In a related study, it is recommended by Chan and Stolfo [40] that data partitions should neither be too large for the time complexity of the learning algorithms nor too small to produce poor classifiers [4]. Given this information, the approach adopted is to randomly select a fixed number of legal examples and merge them with the entire fraud examples. The data partitions are formed by merging all the available X fraud instance (923) with a different set of Y legal instances to form ten X: Y partitions; use sampling with-replacement and without-replacement. The fraud instances: legitimate instance (923:923) with a fraud: legitimate distribution of 50:50; gave the best result and was used throughout the whole experiments in this study. This data partitioning technique is explained in more details in a previously published paper[18].

B. Algorithms Used in the Proposed Models

1) Grading

The term graded is used in the sense of classifications that have been marked as correct or incorrect. The method transforms the classification made by the k different classifiers into k training sets by using the instances k times and attaching them to a new binary class in each occurrence. This class indicates whether the k^{th} classifier yielded a correct or incorrect classification, compared to the real class of the instance.

For each base classifier, one meta-classifier is learned whose task is to classify when the base classifier will misclassify. At classification time, each base classifier classifies the unlabeled instance. The final classification is derived from the classifications of those base classifiers that are classified to be correct by the meta-classification schemes. In case several base classifiers with different classification results are classified as correct, voting, or a combination considering the confidence estimates of the base classifiers, is performed. Grading may be considered as a generalization of cross-validation selection [41], which divides the training data into k subsets, builds k-1 classifiers by dropping one subset at a time and then using it to find a misclassification rate. Finally, the procedure simply chooses the classifier corresponding to the subset with the smallest misclassification. Grading make this decision separately for each and every instance; by using only those classifiers that are predicted to classify that instance correctly. The main difference between grading and combiners (or stacking), are that the former does not change the instance attributes by replacing them with class predictions or class probabilities (or adding them to it). Instead it modifies the class values. Furthermore, in grading several sets of meta-data are created, one for each base classifier.

2) Stacking

Stacking is a technique whose purpose is to achieve the highest generalization accuracy. This method tries to distinguish between reliable classifiers and not reliable. It is used to combine models built by different inducers. The idea is to create a new dataset containing a tuple for each tuple in the original dataset. However, instead of using the original input attributes, it uses the predicted classification of the classifiers as the input attributes. The target attributes remains as in the original training set.

Test instance is first classified by each of the base classifiers. These classifications are fed into a meta-level training set from which a meta-classifier is produced. This classifier combines the different predictions into a final one. It is recommended that the original dataset will be partitioned into two subsets.

The first subset is reserved to form the meta-dataset and the second subset is used to build the base-level classifiers. Consequently the meta-classifier predications reflect the true performance of base-level learning algorithms. Stacking performances could be improved by using output probabilities for every class label from the base-level classifiers. In such cases, the number of input attributes in the meta-dataset is multiplied by the number of classes [21].

3) Voting

In this combining schema, each classifier has the same weight. A classification of an unlabeled instance is performed according to the class that obtains the highest number of votes. Mathematically it can be written as:

$$Class(x) = \underset{c_i \in dom(y)}{\operatorname{argmax}} \sum_{\forall k c_i = \underset{c_j \in dom(y)}{\operatorname{argmax}} \hat{P}_{M_k}(y=c_j|x)} 1 \quad (3)$$

Where M_k denotes classifier k and $\widehat{P_{M_k}}(y = c \mid x)$ denotes the probability of y obtaining the value c given an instance x.

IV. Data Pre-processing and Experimental Setup

A. Data description

In this research the insurance dataset used was originally used by Phua et al. (2004) [42]. This dataset mainly contains automobile insurance claims during the period 1994–1996, comprises 32 variables, with 31 predictor variables and one class variable. It consists of 15,420 samples of which 14,497 are legitimate cases and 923 are fraudulent cases, which mean there are 94% legitimate cases and 6% fraudulent cases.

B. Data Cleaning and Preparation

It is observed that the age attribute in the dataset appeared twice in numerical and categorical form as well; hence the categorical attribute was chosen since it is simpler.

Further, the date of the accident was represented by the four attributes year, month, week of the month and day and date of the insurance claim was represented by three attribute. Thus, a new attribute gap was derived from all these attributes; it represents the time difference between the accident occurrence and insurance claim. Hence, 15,420 examples with 24 predictor variables and one class variable formed the final dataset.

C. Experimental setup

The experiment setup of this paper can be divided into several steps mentioned below:

1) The original dataset is resample using Partitioning-undersampling technique with fraud: legal ratio of 50: 50 and using sampling with-replacement; which was shown to give the best results.
2) The best base-classifiers models from a previous paper; IFDDT, IFDSVM and IFDANN that were designed using Decision Tree, Supper Vector Machine and Artificial Neural Network will be used in this experiment.
3) In hope to enhance the previously proposed IFD models, ensemble combination classifiers were applied to base-classifier IFD models and all possible combinations of different base-classifiers IFD models. Stacking, Grading and Voting; the ensemble combining classifiers were used in this research to form several IFD ensemble modes.
4) Since the dataset was imbalanced the best evaluation and validation measurements are Recall, Precision and the area under the Precision-Recall (PR) curve. Those evaluation and validation measures were explained in more details in previously published paper [18], justifying their usage. A comparison is done based on the evaluation and validation measures results of the proposed IFD ensemble models.
5) For further evaluation of the proposed models the p-value is calculate to test if the difference between the models is statistically significant.
6) The proposed models were applied on another imbalance dataset (German Dataset) are the results of both datasets was compared and analysis.
7) Then the novel proposed IFD models using German dataset were compared with previously designed models using the same dataset.

V. Results and Analysis

Ensemble combining classifiers are applied on base-classifier models to design proposed IFD ensemble models. The experiment is designed to use the best base-classifiers models from a previous paper; IFDDT, IFDSVM and IFDANN that we designed using Decision Tree, Supper Vector Machine and Artificial Neural Network. Three different ensemble combining classifiers are used namely Grading, Stacking and Vote.

A. IFD Models Using Grading

Grading is a meta combining ensemble classifier, that grade the base classifiers [43], as explained in (section III). All possible combinations of the three best classifiers from the previous paper, are formed by using grading also single classifiers were used since it could be applied on single classifier. As mentioned previously the models are evaluated using recall, precision and PRC area, which were recorded. The models are chosen according to the recall value if two or more have the same recall value then they are chose according to precision or both.

The classifier has several parameters, which were adjusted by trial and error choosing the best. The meta-classifier used for the grading is OneR, which was chosen from several chooses. OneR is a classifier that uses the minimum-error attribute for prediction of numeric attributes [44].

The ten partitions of the subsample are used to design an ensemble model and then an average is calculated for the ten partitions. The models of the ten partitions of the subsample are combined in one model using averaging. The seven ensemble models that were designed using grading are compared using recall, precision and PRC area as shown in Table (1).

Grading improved the IFDANN model from 87.7% to 89.5% but the IFDSVM and IFDDT did not improve as their recall value decreased. IFDSVM model using grading, recall decreased to 93.4% from 94.1%. IFDDT model using grading; recall decrease to 94.3% from 95.8%. The best model designed using grading is the ensemble combination of IFDSVM and IFDDT; which recall is 95.4%.

Grading	Recall	Precision	PRC Area

SVM	93.6%	69.0%	70.4%
ANN	89.5%	69.3%	69.3%
DT	94.3%	69.1%	70.9%
SVM-ANN	93.6%	69.1%	70.6%
SVM-DT	95.4%	69.1%	71.4%
ANN-DT	92.9%	69.1%	70.4%
SVM-ANN-DT	94.3%	69.2%	70.9%

Table 1: Grading of base-classifier models combination.

B. IFD Models Using Stacking

Another IFD models were designed using stacking, which is a meta combining ensemble classifier [45]. Stacking meta-classifier has several parameters; whose values were chosen by trial and error. This classifier requires a meta-classifier; to be a numeric prediction scheme. The meta-classifier chose is Linear Regression which was chosen after testing a lot of meta-classifiers. This classifier learns a Simple Linear Regression model. It picks the attribute that result in the lowest squared error. Missing values are not allowed. It can only deal with numeric attributes.

The best base-classifiers models and all possible combinations of those models were used to design the ensemble IFD models. After combing the ten partitions of the subsample using averaging the resulting models are compared according to recall, precision ad PRC area as shown in Table (2). Applying stacking on the single base-classifier model improved the IFDANN model from 87.7% to 94.3% but the other two models did not improve. IFDSVM model using stacking, recall decreased to 93.6% from 94.1%. IFDDT model using stacking; recall decrease to 94.4% from 95.8%. The best model designed using stacking is the ensemble combination of IFDANN and IFDDT; which recall is 94.4%.

STACKING	Recall	Precision	PRC Area
SVM	93.6%	69.0%	71.8%
ANN	94.3%	68.9%	77.7%
DT	94.4%	69.2%	75.9%
SVM-ANN	94.1%	69.1%	77.9%
SVM-DT	94.1%	69.1%	76.7%
ANN-DT	94.6%	69.1%	78.7%
SVM-ANN-DT	94.0%	69.2%	78.6%

Table 2 : Stacking of base-classifier models combination

C. IFD models using Vote

In this Section, more IFD models were designed using voting, which is a simple combining ensemble classifier. It is a class for combining classifiers; different combinations of probability estimates for classification are available [46, 47] . The classifier has several parameters whose best values were found by testing several values by trial and error. The best base-classifiers models and all possible combinations of those models were used to design the ensemble IFD models. After combing the ten partitions of the subsample using averaging the resulting models are compared according to recall, precision ad PRC area as shown in Table (3). Applying vote on the single base-classifier model improved the IFDANN model a lot from 87.7% to 92.6% but the other two models didn't improve. IFDSVM model using voting, recall decreased to 93.6% from 94.1%. IFDDT model using voting; recall decrease to 94.3% from 95.8%. The best model designed using voting is the IFDDT; which recall is 94.3%

VOTE	Recall	Precision	PRC Area
SVM	93.6%	69.0%	70.5%
ANN	92.6%	69.2%	76.2%
DT	94.3%	69.1%	75.1%
SVM-ANN	93.6%	69.0%	78.3%
SVM-DT	93.8%	69.0%	76.0%
ANN-DT	93.5%	69.2%	79.2%
SVM-ANN-DT	93.8%	69.3%	79.2%

Table 3 : Vote of base-classifier models combination

D. Evaluating the IFD Ensemble Models

For evaluating the proposed IFD models, two methods were followed. First proving that the different in the recall between the models is statistically significant difference and then applying those models on another imbalance dataset and comparing between the results.

The p-values for the designed IFD models were calculated, using t-test with nil-hypothesis that the mean of samples are equal. All models are compared with the best model DT. The models have p-value less than 0.05; thus the nil-hypothesis is reject; meaning that the difference between the models is a significant difference [48]. All designed IFD models are shown in Table (4) sorted according to their Recall and also showing their p-value.

The Recall of all designed models shown in Table (4), show that the application of ensemble combining classifiers produced a strong classifiers but still IFDDT with recall 95.8%, is the best designed model. Next to IFDDT model came model using grading (SVM-DT) (IFD-G-SVM-DT) with recall 95.4%, model using stacking (ANN-DT) (IFD-S-ANN-DT) with recall 94.6% and the model using grading(SVM-ANN-DT) (IFD-G-SVM-ANN-DT) with recall 94.3%.

The proposed models were applied on German dataset to use for evaluation of the proposed models. A comparing between averaged partitions and German dataset models is shown in Table (5). The results of applying the proposed models on German dataset, is shown on Table (5), the best two models that gave best results are the model using Grading

(SVM-ANN-DT) with recall 78.4% and the model using Vote (SVM-ANN-DT) with recall 77.3%. This results show that the ensemble combing classifiers give enhance the IFD models that were designed at first using base-classifiers.

	Recall %	Precision	PRC Area	p-value
DT	95.8	69.4%	73.0%	
grading(SVM-DT)	95.4	69.1%	71.4%	1.325E-11
stacking(ANN-DT)	94.6	69.1%	78.7%	6.787E-22
stacking(DT)	94.4	69.2%	75.9%	1.555E-30
grading(SVM-ANN-DT)	94.3	69.2%	70.9%	2.342E-09
vote(DT)	94.3	69.1%	75.1%	1.000E+00
grading(DT)	94.3	69.1%	70.9%	1.043E-15
stacking(ANN)	94.3	68.9%	77.7%	7.866E-82
stacking(SVM-ANN)	94.1	69.1%	77.9%	2.408E-28
stacking(SVM-DT)	94.1	69.1%	76.7%	1.487E-19
SVM	94.1	68.9%	70.7%	1.023E-21
stacking(SVM-ANN-DT	94.0	69.2%	78.6%	1.364E-18
vote(SVM-ANN-DT)	93.8	69.3%	79.2%	3.418E-06
vote(SVM-DT)	93.8	69.0%	76.0%	1.425E-08
grading(SVM-ANN)	93.6	69.1%	70.6%	2.531E-09
vote(SVM-ANN)	93.6	69.0%	78.3%	1.372E-10
stacking(SVM)	93.6	69.0%	71.8%	5.047E-42
vote(SVM)	93.6	69.0%	70.5%	3.267E-12
grading(SVM)	93.6	69.0%	70.4%	2.473E-10
vote(ANN-DT)	93.5	69.2%	79.2%	8.324E-25
grading(ANN-DT)	92.9	69.1%	70.4%	2.352E-14
vote(ANN)	92.6	69.2%	76.2%	6.543E-32
ANN	89.7	69.0%	76.7%	6.788E-33
grading(ANN)	89.5	69.3%	69.3%	2.134E-13

Table 4: IFD base-classifiers and ensemble models p-value.

	Averaged Partitions Recall	German Dataset Recall
Grading (SVM-ANN-DT)	94.3%	78.4%
Vote (SVM-ANN-DT)	93.8%	77.3%
SVM	94.1%	77.0%
Vote (SVM)	93.6%	77.0%
Vote (SVM-ANN)	93.6%	77.0%
Stacking (SVM)	93.6%	77.0%
Stacking (SVM-ANN)	94.1%	77.0%
Stacking (SVM-ANN-DT)	94.0%	77.0%
Grading (ANN-DT)	92.9%	76.9%
ANN	89.7%	76.8%
Stacking (SVM-DT)	94.1%	76.4%
Grading (SVM-ANN)	93.6%	76.0%
Grading (SVM-DT)	95.4%	75.8%
Grading (ANN)	89.5%	75.5%
Vote (ANN)	92.6%	75.3%
Vote (SVM-DT)	93.8%	75.0%
Grading (DT)	94.3%	74.8%
Stacking (DT)	94.4%	74.7%
Stacking (ANN-DT)	94.6%	74.7%
Stacking (ANN)	94.3%	74.4%
Vote (ANN-DT)	93.5%	74.3%
DT	95.8%	73.5%
Vote (DT)	94.3%	73.5%
Grading (SVM)	93.6%	60.2%

Table 5 : comparing between averaged partitions and German dataset models

VI. Conclusions

Ensemble combining classifiers were applied on IFD models that were designed in previous paper using an imbalance automobile insurance fraud detection dataset. The imbalance dataset problem was solved by a novel proposed technique "partitioning-undersampling".

The proposed models were evaluated according to their recall. The model with the highest recall is IFDDT which was designed using decision tree base classifier, but the second, third and fourth best classifiers are; model using grading

(SVM-DT) (IFD-G-SVM-DT) with recall 95.4%, model using stacking (ANN-DT) (IFD-S-ANN-DT) with recall 94.6% and the model using grading (SVM-ANN-DT) (IFD-G-SVM-ANN-DT) with recall 94.3%; which all were designed using ensemble classifiers.

These novel proposed models were applied on another imbalance dataset again the models with the highest recall were the ensemble combination of the three base-classifiers models. The best two models are the model using Grading (SVM-ANN-DT) with recall 78.4% and the model using Vote (SVM-ANN-DT) with recall 77.3%, this proves that ensemble combining classifiers produce powerful IFD models. The difference between those models was shown to be statistically significant difference by calculating their p-value.

References

[1] J. Pearsall, *The concise Oxford dictionary-/ed. by Judy Pearsall*: Oxford [etc.]: Oxford University Press, 1999.

[2] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119,* 2010.

[3] J.-H. Wang, Y.-L. Liao, T.-m. Tsai, and G. Hung, "Technology-based Financial Frauds in Taiwan: Issues and Approaches," in *SMC*, 2006, pp. 1120-1124.

[4] C. Phua, Damminda Alahakoon, and Vincent Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter* vol. 6, pp. 50-59, 2004.

[5] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *Evolutionary Computation, IEEE Transactions on,* vol. 1, pp. 67-82, 1997.

[6] J. M. Pérez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and J. I. Martín, "Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance," in *Pattern Recognition and Data Mining*, ed: Springer, 2005, pp. 381-389.

[7] M. Farquad, V. Ravi, and S. B. Raju, "Analytical CRM in banking and finance using SVM: a modified active learning–based rule extraction approach," *International Journal of Electronic Customer Relationship Management,* vol. 6, pp. 48-73, 2012.

[8] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Engineering Applications of Artificial Intelligence,* vol. 37, pp. 368-377, 2015.

[9] I. Ibarguren, J. M. Pérez, J. Muguerza, I. Gurrutxaga, and O. Arbelaitz, "Coverage based resampling: Building robust consolidated decision trees," *Knowledge-Based Systems,* 2015.

[10] C. Elkan, "Magical thinking in data mining: lessons from CoIL challenge 2000," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 426-431.

[11] N. Lavrač, H. Motoda, T. Fawcett, R. Holte, P. Langley, and P. Adriaans, "Introduction: Lessons learned from data mining applications and collaborative problem solving," *Machine learning,* vol. 57, pp. 13-34, 2004.

[12] I. Bose and R. K. Mahapatra, "Business data mining—a machine learning perspective," *Information & management,* vol. 39, pp. 211-225, 2001.

[13] E. Turban, R. Sharda, D. Delen, and T. Efraim, *Decision support and business intelligence systems*: Pearson Education India, 2007.

[14] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI magazine,* vol. 13, p. 57, 1992.

[15] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Networking, sensing and control, 2004 IEEE international conference on*, 2004, pp. 749-754.

[16] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems,* vol. 50, pp. 559-569, 2011.

[17] A. K. I. Hassan and A. Abraham, "Computational Intelligence Models for Insurance Fraud Detection: A Review of a Decade of Research," *Journal of Network and Innovative Computing,* vol. 1, pp. 341-347, 2013.

[18] A. K. I. Hassan and A. Abraham, "Modeling Insurance Fraud Detection Using Imbalanced Data Classification," in *Advances in Nature and Biologically Inspired Computing*, ed: Springer, 2016, pp. 117-127.

[19] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *Power Systems, IEEE Transactions on,* vol. 21, pp. 53-60, 2006.

[20] Y. Peng, G. Kou, A. Sabatka, Z. Chen, D. Khazanchi, and Y. Shi, "Application of Clustering Methods to Health Insurance Fraud Detection," in *Service Systems and Service Management, 2006 International Conference on*, 2006, pp. 116-120.

[21] L. Rokach, "Ensemble Methods for Classifiers," in *Data Mining and Knowledge Discovery Handbook*, ed: Springer US, 2010, pp. 957-980.

[22] M. Sternberg and R. G. Reynolds, "Using cultural algorithms to support re-engineering of rule-based expert systems in dynamic performance environments: a case study in fraud detection," *Evolutionary Computation, IEEE Transactions on,* vol. 1, pp. 225-243, 1997.

[23] P. L. Brockett, X. Xia, and R. A. Derrig, "Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud," *Journal of Risk and Insurance,* pp. 245-274, 1998.

[24] S. Tennyson and P. Salsas-Forn, "Claims auditing in automobile insurance: fraud detection and deterrence objectives," *Journal of Risk and Insurance,* vol. 69, pp. 289-308, 2002.

[25] M. Artís, M. Ayuso, and M. Guillén, "Modelling different types of automobile insurance fraud behaviour in the Spanish market," *Insurance: Mathematics and Economics,* vol. 24, pp. 67-81, 1999.

[26] M. Artís, M. Ayuso, and M. Guillén, "Detection of automobile insurance fraud with discrete choice models and misclassified claims," *Journal of Risk and Insurance,* vol. 69, pp. 325-340, 2002.

[27] S. B. Caudill, M. Ayuso, and M. Guillen, "Fraud detection using a multinomial logit model with missing information," *Journal of Risk and Insurance,* vol. 72, pp. 539-550, 2005.

[28] E. B. Belhadji, G. Dionne, and F. Tarkhani, "A model for the detection of insurance fraud," *Geneva Papers on Risk and Insurance. Issues and Practice,* pp. 517-538, 2000.

[29] J. Pinquet, M. Ayuso, and M. Guillen, "Selection bias and auditing policies for insurance claims," *Journal of Risk and Insurance*, vol. 74, pp. 425-440, 2007.
[30] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A Comparison of State - of - the - Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection," *Journal of Risk and Insurance*, vol. 69, pp. 373-421, 2002.
[31] S. Viaene, R. A. Derrig, and G. Dedene, "A case study of applying boosting Naive Bayes to claim fraud diagnosis," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, pp. 612-620, 2004.
[32] S. Viaene, G. Dedene, and R. A. Derrig, "Auto claim fraud detection using Bayesian learning neural networks," *Expert Systems with Applications*, vol. 29, pp. 653-666, 2005.
[33] W. Xu, S. Wang, D. Zhang, and B. Yang, "Random Rough Subspace Based Neural Network Ensemble for Insurance Fraud Detection," in *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*, 2011, pp. 1276-1280.
[34] M. Vasu and V. Ravi, "A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance," *International Journal of Data Mining, Modelling and Management*, vol. 3, pp. 75-105, 2011.
[35] S. Viaene, M. Ayuso, M. Guillen, D. Van Gheel, and G. Dedene, "Strategies for detecting fraudulent claims in the automobile insurance industry," *European Journal of Operational Research*, vol. 176, pp. 565-583, 2007.
[36] R. Bhowmik, "Detecting auto insurance fraud by data mining techniques," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 2, pp. 156-162, 2011.
[37] P. L. Brockett, R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert, "Fraud classification using principal component analysis of RIDITs," *Journal of Risk and Insurance*, vol. 69, pp. 341-371, 2002.
[38] J. Pathak, N. Vidyarthi, and S. L. Summers, "A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims," *Managerial Auditing Journal*, vol. 20, pp. 632-644, 2005.
[39] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *Intelligent Systems and their Applications, IEEE*, vol. 14, pp. 67-74, 1999.
[40] P. K. Chan and S. J. Stolfo, "A comparative evaluation of voting and meta-learning on partitioned data," in *ICML*, 1995, pp. 90-98.
[41] C. Schaffer, "Selecting a classification method by cross-validation," *Machine learning*, vol. 13, pp. 135-143, 1993.
[42] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, pp. 50-59, 2004.
[43] A. K. Seewald and J. Fürnkranz, "An evaluation of grading classifiers," in *Advances in Intelligent Data Analysis*, ed: Springer, 2001, pp. 115-124.
[44] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine learning*, vol. 11, pp. 63-90, 1993.
[45] A. K. Seewald, "How to make stacking better and faster while also taking care of an unknown weakness," in *Proceedings of the nineteenth international conference on machine learning*, 2002, pp. 554-561.
[46] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms: John Wiley & Sons, 2004.
[47] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 226-239, 1998.
[48] P. J. Bickel and K. A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, volume I vol. 117: CRC Press, 2015.