# Fraud Detection Supervised Machine Learning Models for an Automobile Insurance

Nikhil Rai, Pallav Kumar Baruah, Satya Sai Mudigonda and Phani Krishna Kandala

**Abstract**— In this paper, we have built a robust fraud detection model, built upon an existing fraud detection research.Usually, machine learning models do not perform well in the presence of class-imbalance in the dataset. They tend to favor the majority class where the main objective was to detect minority class. We have used one such oversampling-technique MWMOTE[1] to handle this class imbalance problem and build three different models: Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF). We found that our proposed method is giving us good results in comparison to the existing methods on the automobile insurance dataset, "carclaims.txt."

**Index Terms**— Insurance Company, Fraud, Fraud Detection, Class-Imbalance, over-sampling, machine learning models.

## 1 INTRODUCTION

There has been a rapid growth in the insurance industry with respect to a large amount of data. As the data size increases, the traditional approach is finding tough to work on it and becoming a tedious job to identify the fraudulent claims. An insurance company, by its nature, is very susceptible to fraud. Insurance companies are losing a huge amount of money in such fraudulent claims. Once such industry is the automobile insurance company.


- Nikhil Rai is currently pursuing Master degree in Computer Science in Department of Mathematics and Computer Science in Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India. PH: +91 6295877384, Email: poorna.sravan@gmail.com.
- Pallav Kumar Baruah, Head of Department, Department of Mathematics and Computer Science in Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India. PH: +91 9440699887, Email: pkbaruah@sssihl.edu.in.
- Satya Sai Mudigonda is a professionally qualified associate actuary and management consultant. He is currently teaching the postgraduate students in Department of Mathematics and Computer Science in Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India. PH: +91 9603573032, Email:satyasaibabamudigonda@sssihl.edu.in
- Phani Krishna Kandala, is currently Assistant Vice President in Swiss Re. He has done in Master's from Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India. PH:+91 91 82 472136, Email:kandala.phanikrishna@gmail.com.


Generally, an automobile insurance contract is signed between an insurance company, which is also called an insurer, and a customer also called an insured. In basic terms, it is a contract between an insurer and an insured which provides a financial support to an insured by an insurer during the case of vehicular theft or damage. Fraud in an insurance can be broadly classified into two categories:

- Hard Fraud: A hard fraud is a type of fraud which requires scheming, planning and sometimes even someone from the inside to obtain the financial benefits from an insurance company. It can be attributed to premeditated, planned and deliberate.
- Soft Fraud: Soft fraud is a more prevalent form of fraud and also known as opportunistic fraud.

Automobile insurance fraud occurs by getting into an accident on purpose. It also occurs when fake documents are submitted regarding casualties in a staged accident. The main motive behind this is to get the financial benefits that insurer have promised while taking an insurance policy (Ngai et al., (2011)[2]). According to the Federal Bureau of Investigation (FBI) insurance fraud division, "The total cost of insurance fraud (non-health insurance) is estimated to be more than $40 billion per year. The FBI states that 4% of the money that the insurance industry makes is lost to insurance fraud. The Association of British Insurers (ABI) investigated the increase in the number of false claims and found that it was 18% more than the previous year (Cutting corners, (2015)[3]). According to the National Insurance Crime Bureau (NICB), questionable claims continue to rise each year, with a 34 percent increase between 2008 and 2011. It is reported in [4], that approximately 21% – 36% auto-insurance claims contain elements of suspected fraud but only less than 3% of the suspected fraud is prosecuted.

All these fraud statistics show the importance of handling the fraudulent claims and help the firms in incurring a huge amount of losses. But insurance fraud detection is a challenging problem. Traditional fraud detection methods are heavily dependent on auditing and expert inspection. These methods are costly and inefficient. It requires money and time. On the other hand, fraud needs to be detected prior to the claim payment. Since data mining and machine learning techniques have huge potential in analyzing a large amount of data and detecting the suspicious and fraudulent claims in a timely manner, these can be used to build a model to identify the fraudulent claims.

One of the main problems with these machine learning models is that they suffer from the problem of class imbalance in the dataset on which these models have been built. A class-imbalance problem occurs in the data when the total number of one class (minority) is far less than the total number of another class (majority). In such a case, learning becomes difficult for the models. Most of the time, models tend to favor the majority class. Learning from the imbalanced dataset is itself another research area. In this paper we have used one such over-sampling technique MWMOTE[1] for handling this problem and build three different models: Support Vector Machine

(SVM), Decision Tree (DT) and Random Forest (RF). We found that our proposed method of fraud detection for automobile insurance fraud detection is giving us good results compare to the existing state of the art.

The paper is organized as follows: Section 2 talks about the literature survey in the area of fraud detection and class-imbalance problem. Section 3 talks about the algorithm used in the proposed approach. In Section 4, we talk about the proposed method and results on the automobile insurance dataset, "carclaims.txt." Finally, Section 5 presents the conclusion and future work.

## 2 LITERATURE REVIEW

In this section, we present the previous works in the area of fraud detection and the techniques used for solving imbalance dataset problems. The literature review is divided into two parts: insurance fraud detection and techniques for handling the imbalanced problem in a dataset.

### 2.1 Insurance Fraud Detection

The use of data analytics and data mining is changing the insurance fraud detection method. Data mining deals with the finding of information and hidden patterns which are statistically reliable, unknown previously and actionable from data [5]. In [6], data mining is defined as a method of finding useful patterns in data that can be helpful while making a decision. A meta-learning system [7] is developed for detecting the fraud. This system combines the result of different local based models at different sites and comes up with more accurate fraud detection tools. (Chan et al. (1999)[8] and Stolfo et al. (2000)[9]) extended this work and developed a data mining model which is distributed and scalable. It was used for evaluating the classification techniques. (Brockettetal.(2002)[10])proposed a mathematical method for an apriori classification of objects when no training data with target sample exits. They used RIDIT scores and found that an insurance fraud detector can increase the chances of targeting the appropriate claims and reduce uncertainty. (Phua et al. (2004)[11]) proposed a hybridization of two techniques stacking and bagging meta-classifiers. They introduced a fraud detection method which makes use of a single meta-classifier (stacking) to choose the best base classifiers, and then combine these base classifiers prediction (bagging) to improve cost savings. (Viaene et al. (2005)[12]) used a Bayesian learning neural networks for auto claim fraud detection. The use of an automatic relevance determination objective function scheme determines which inputs are most informative to the trained neural network model. (Pathak et al. (2005)[13]) used the fuzzy logic concept for finding the illegitimate claims from a bunch of settled insurance claims. (Bermudez et al. (2008)[14]) introduced asymmetric Bayesian dichotomous logit model for finding the fraudulent insurance claims found in a Spanish insurance market. They have developed this model using data augmentation and Gibbs sampling and found out that the use of an asymmetric or skewed logit link significantly improves the percentage of cases that are correctly classified after the model estimation. (Sublej et al. (2011)[15]) used a graph-based social network model in order to identify frauds in automobile insurance. They have developed an Iterative Assessment Algorithm (IAA) that was based on Graph Components for identifying the suspicious claims. Each point in the graph is given a suspicion score and then suspicious claim is determined by analyzing the edges present within their neighboring nodes. (Xu et al (2011)[16]) used a random rough subspace based neural network ensemble for insurance fraud detection. They have divided the whole dataset into the training set and testing set and the training set is divided further into multiple training subsets by selecting r-dimensional random subspaces. Different classifiers are trained on this multiple training sets to build a trained model. In the end, a final decision is taken by taking a majority voting of each model. (Sundarkumar and Ravi (2015)[17]) used a One Class Support Vector Machine (OCSVM) as an under-sampling technique to handle the class-imbalance problem and five different classifiers are trained from the balanced dataset and found that Decision Tree was giving the best result compared to all four different classifiers. (Nian et al.(2016)[18]) proved an unsupervised model auto insurance fraud detection. They have used an unsupervised spectral ranking for anomaly and found their method was surpassing the existing outlier-based fraud detection model. (Subudhi et al. (2017)[19]) proposed the use of fuzzy cmean clustering for making the dataset balanced and used the thresholding technique to identify whether the majority samples are outlier or not.

### 2.2 Technique for handling class-imbalance dataset problem

In the presence of imbalanced dataset, machine learning models tend to favor the majority class, where model's performance is not good for the minority class [20] [21]. This happens because machine learning models will try to return the most correct predictions depending on the entire dataset, which results in them classifying all the data as belonging to the larger class. This larger class is of least interest to the data-mining problem where the main goal is to identify the minority class. For example, the main goal in insurance fraud detection is to identify the fraudulent (minority) data, not the non-fraudulent (majority) data. In this section, we review the past work reported in different techniques in handling this problem.

(Hart (1968)[22]) proposed an under-sampling method, Condensed Nearest Neighbor (CNN). This method initially starts with two blank datasets A and B. Then randomly a sample is drawn and placed it in dataset A, while the rest of the samples are placed in dataset B. Then one instance from dataset B is scanned by using the dataset A as the training set. If an instance in B is misclassified, it is transferred from B to A. The process repeats until no instances are transferred from B to A. (Sternberg and Reynolds (1997)[23]) solved the problem by searching manually for the features that cause type 1 error (false positive) and type 2 error (false negative) and use these features to design the model. (Laurikkala (2001)[24]) proposed Neighborhood Cleaning Rule (NCR). It uses Wilson's Edited

Nearest Rule to remove selected majority class examples. Other techniques involve the generation of new synthetic samples from the minority samples. (Chawla et al. (2002)[25]) proposed Synthetic Minority Oversampling Technique (SMOTE) approach, where the new synthetic minority samples are generated rather than just oversampling with replacement. (He et al. (2008)[26]) proposed the Adaptive Synthetic (ADASYN) oversampling technique which was an improved version of SMOTE. It does same as the SMOTE just with a minor improvement. After creating those new synthetic minority samples, it adds a random small value to these thus making it more realistic. (Han et al. (2005)[27]) proposed the new oversampling technique to handle the borderline minority samples. Borderline samples are those samples that are close to the decision boundary. These samples are the ones that are most likely to be miss-classified. They proposed the used of δ to identify the minority samples as borderline samples. In some of the situation the mentioned oversampling techniques do not work, (Baruaetal.(2014)[1]) proposed an over-sampling technique, MWOTE.

(SundarKumar et al. (2015)[28]) proposed the use of k-reversed Nearest Neighborhood and One Class support vector machine (OCSVM) as an under-sampling technique for handling the class-imbalance problem. (Subudhi et al. (2017)[19]) proposed the use of fuzzy c-means algorithm to identify the majority samples as an outlier and used it as an under-sampling technique. (Sudarsun Santhiappan et al. (2018)[29]) proposed TODUS, a top-down oriented directed under-sampling algorithm that follows the estimated data distribution to draw samples from the dataset. (Douzas et al. (2018)[30]) proposed the use of Conditional Generative Adversarial Networks to approximate the true data distribution and generate data for the minority class of various imbalanced datasets.

## 3 Algorithm Used In The Proposed Approach

In this section, we will discuss an algorithm, Majority Weighted Minority Oversampling Technique (MWMOTE), which is used for handling the class-imbalance problem.

### 3.1 Majority Weighted Minority Oversampling Technique (MWMOTE)

Majority Weighted Minority Oversampling Technique is introduced in [1]. It is one of the oversampling technique to handle the class imbalance problem present in the dataset. It generates new synthetic samples from seed samples.

Oversampling methods like Synthetic Minority Oversampling Technique (SMOTE)[25], Borderline Smote (BrdSMOTE)[27], Adaptive Synthetic Minority Techniques (ADYSN)[26] fail to identify the borderline samples in some situation. Borderline samples are those samples which lie closer to the decision boundary. These samples are the one which can be miss-classified by the classifier. MWMOTE tries to handle this situation and identifies the borderline samples by assigning a weight to the hard-to learn minority samples based on the majority samples. This method focusses on two objectives: one is to improve the sample selection scheme and another one is to improve the scheme of generation of synthetic samples based on the hard-to-learn samples. It comprises mainly three important main stages:

- Firstly, samples which are hard-to-learn and the most important minority samples are identified.
- Secondly, each of the hard-to-learn minority samples is given weight based on its importance in the data. These weights are based on the majority samples.
- Lastly, new synthetic minority samples are generated following a similar strategy to SMOTE.

One can find the full algorithm for the MWMOTE in [1].

## 4 Proposed Method and Result

This section talks about an approach for building a fraud detection model for identifying the fraudulent claims in the automobile dataset.
In order to identify the fraudulent claims, we have proposed a new approach which is shown in Figure 1:

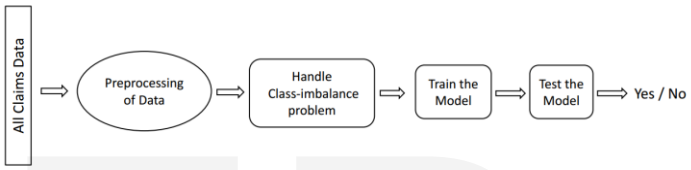


Figure 1

First our model focus on the preprocessing of data, which is an important step for building a good classifier. Details of preprocessing steps are given in the following section. Next, the model handles the problem of imbalanced data. In order to tackle this problem, we have used an above mentioned oversampling technique MWMOTE. With the help of this technique, we have synthetically generated minority samples. After handling the class-imbalance problem, we have built three different classifiers: Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF). We have used a 10-fold cross validation for training and testing these classifiers and compared our result with the existing state of art method present in the literature. The results of our method and its comparison are shown below in the following sections.

### 4.1 Data Description

"carclaims.txt" dataset is the only publicly available automobile insurance dataset and is taken from (Phua et al. (2004)[11]). This dataset is provided by Angoss KnowledgeSeeker Software. It consists of 15420 claim instances from January 1994 to December 1996, having 14,497 genuine samples (94%) and 923 fraud instances (6%). Hence the dataset is highly imbalanced. The dataset has 6 ordinal features and 25 categorical attributes. The description of each of the attributes is shown in Figure 2.

### 4.2 Data Preprocessing

In this paper, we have used one-hot encoding and binary encoding representation for representing the categorical attributes present in the dataset. Some of the procedures of data preprocessing are taken from (Phua et al. (2004)[11]). Once

| S. No | Attribute name | Description |
|---|---|---|
| 1 | Month | Month in which accident took place |
| 2 | Week of month | Accident week of month |
| 3 | Day of week | Accident day of week |
| 4 | Month claimed | Claim month |
| 5 | Week of month claimed | Claim week of month |
| 6 | Day of week claimed | Claim day of week |
| 7 | Year | 1994, 1995 and 1996 |
| 8 | Make | Manufacturer of the car (19 companies) |
| 9 | Accident area | Rural or urban |
| 10 | Gender | Male or female |
| 11 | Marital status | Single, married, widow and divorced |
| 12 | Age | Age of policy holder |
| 13 | Fault | Policy holder or third party |
| 14 | Policy type | Type of the policy (1–9) |
| 15 | Vehicle category | Sedan, sport or utility |
| 16 | Vehicle price | Price of the vehicle with 6 categories |
| 17 | Rep. number | ID of the person who process the claim(16 ID's) |
| 18 | Deductible | Amount to be deducted before claim disbursement |
| 19 | Driver rating | Driving experience with 4 categories |
| 20 | Days: policy accident | Days left in policy when accident happened |
| 21 | Days: policy claim | Days left in policy when claim was filed |
| 22 | Past number of claims | Past number of claims |
| 23 | Age of vehicle | Vehicle's age with 8 categories |
| 24 | Age of policy holder | Policy holder's age with 9 categories |
| 25 | Policy report filed | Yes or no |
| 26 | Witness presented | Yes or no |
| 27 | Agent type | Internal or external |
| 28 | Number of supplements | Number of supplements |
| 29 | Address change claim | No of times change of address requested |
| 30 | Number of cars | Number of cars |
| 31 | Base policy(BP) | All perils, collision or liability |
| 32 | Class | Fraud found (yes or no) |

Figure 2

these steps are done, the data normalization procedure is applied to the dataset so that all the features have value in the range [0,1]. Since different ranges of attributes can affect the model's performance by giving importance to high valued attributes, data normalization ensures that every data point will get an equal chance rather than high valued attributes.

### 4.3 Performance Metric

To evaluate how our models are performing, we have used the five standard metrics: Accuracy, Precision, Recall/Sensitivity, Specificity, and F1-Score. These metrics measure the effectiveness and usefulness of the model.

• Accuracy: It is the most common performance measure and is defined as the ratio of correctly predicted observation to the total number of observation. Accuracy is a great measure but only when the given dataset is symmetric and balanced. It is given by

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

• Precision: It is defined as the ratio of correctly predicted positive observation to the total predicted positive observations. High precision relates to the low false positive rate. It is given by

$$\text{Precision} = TP / (TP + FP)$$

• Recall (Sensitivity): It is defined as the number of observations correctly identified as positive out of total true positives. It is given by

$$\text{Recall} = TP / (TP + FN)$$

• Specificity: It is defined as the number of observations correctly identified as negatives out of total negatives: It is given by

$$\text{Specificity} = TN / (TN + FP)$$

• F1-Score: It is defined as the harmonic mean of precision and recall. Therefore, this score takes both the false positive and false negative into account. It is more useful than accuracy especially if the dataset has an uneven class distribution. It is given by

$$F1\text{−}Score = 2 * (Precision * Recall) / (Precision + Recall)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive and false negative respectively. Note than recall expresses the ability to find all the relevant instances in a dataset whereas precision expresses the proportion of the data points our model says was relevant actually were relevant. In the case of imbalanced data, the model with the highest accuracy is not a good model. So, we have chosen the model with the highest recall (sensitivity) as the optimal one since recall identifies the number of fraudulent instances.

### 4.4 Results

This section contains the results of our proposed approach. We have trained and build three different model: Support vector machine (SVM), Decision Tree (DT) and Random Forest (RF). These three different models have been built and tested on the publicly available dataset, "carclaims.txt". We have used 10-fold cross-validation. The following tables contain the result of the three different models.

| Model | Accuracy | Precision | Recall | Specificity | F1-Score |
|---|---|---|---|---|---|
| SVM | 94.00 | 68.38 | 3.79 | 99.89 | 7.18 |
| Decision Tree | 88.83 | 15.95 | 20.25 | 93.20 | 17.83 |
| Random Forest | 94.09 | 66.50 | 2.49 | 99.92 | 4.78 |

Table 1: Results of models without applying MWMOTE

From Table 1, we can see the all the three different models have good accuracy. Compared to all the models, Random Forest has the highest accuracy. Based on the accuracy only, we cannot say that Random Forest is the best model compared to the other two. Its' recall (sensitivity) is only 2.49%, which means that almost all the fraudulent observation are being miss-classified by the model. Hence we cannot say that it is the best model. Not only for the Random Forest, recall for the other two models is also not acceptable. All the three models have low recall even though their accuracy is more. This clearly shows that the dataset, "carclaims.txt" is not symmetric and has an imbalanced data problem. Table 2 shows the result of our method after handling this imbalanced data problem.

| Model | Accuracy | Precision | Recall | Specificity | F1-Score |
|---|---|---|---|---|---|
| SVM | 77.50 | 71.65 | 91.03 | 63.98 | 80.18 |
| Decision Tree | 96.97 | 94.33 | 99.97 | 93.95 | 97.07 |
| Random Forest | 99.64 | 99.32 | 99.97 | 99.32 | 99.64 |

Table 2: Result of models after applying MWMOTE

We have applied the technique called MWMOTE, for handling the imbalanced problem. From Table 2, we found that all three models' recall has increased and models are now able to identify the fraudulent cases more properly. Almost all the models' performance metrics have increased except for the SVM. We found that SVM's accuracy has decreased. This can be due to the fact that the new synthetic samples generated by the MWMOTE are over-lapping with others samples. Hence, SVM is wrongly predicting the non-fraudulent cases. In our proposed approach, considering the highest accuracy and recall, we found that Random Forest is the optimal model and giving the best result compared to the other two models. The Receiver Operating Characteristic (ROC) for the above models are shown in Figure 3, 4 and 5.

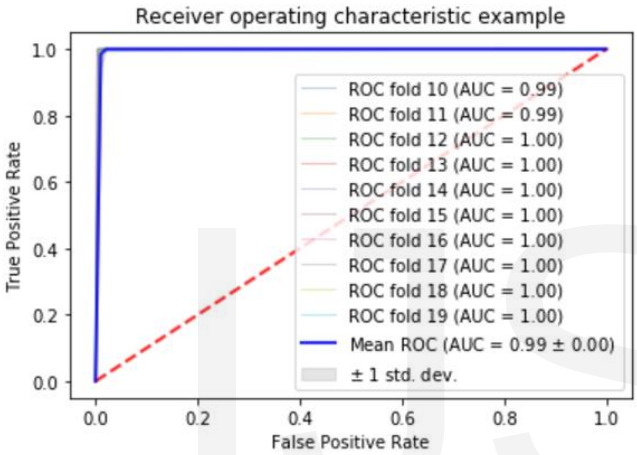


Figure 3: ROC of Random Forest for 10-Folds

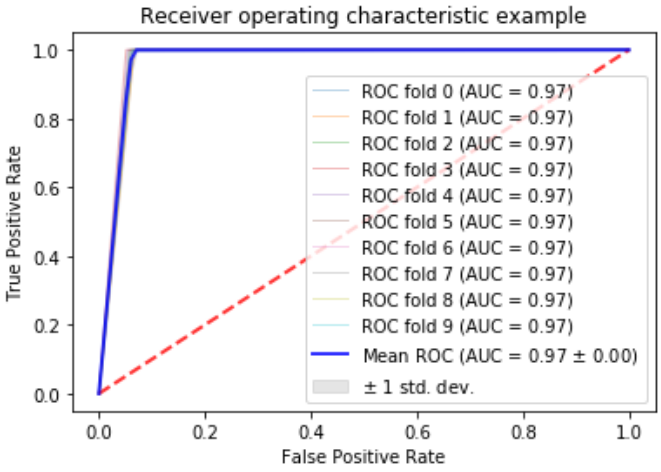


Figure 4: ROC of Decision Tree for 10-Folds

From the figures of ROC also, we could make out that the

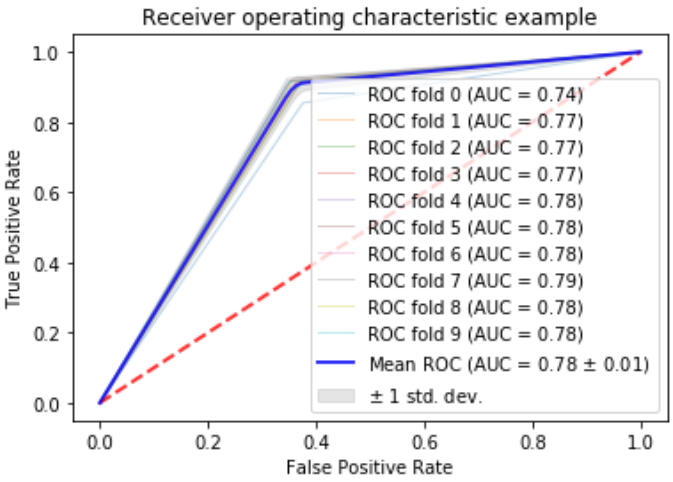


Figure 5: ROC of SVM for 10-Folds

mean Area Under the Curve (AUC) is maximum for Random Forest. This also supports our above claim of Random Forest is the best model compared to Decision Tree and Support Vector

#### 4.4.1 Comparison of our results with the existing results in academic literature

Since this dataset, "carclaims.txt", is publicly available, various researchers have successfully used and have built the classifiers based on this dataset. They have used it for exhibiting their proposed system's performance. Some of the papers that have used this dataset are (Xu et al. (2010)[16]), (Sundarkumar et al. (2015)[28]), (Sundarkumar and Ravi (2015)[17]), (Nian et al. (2016)[18]) and (Subudhi and Panigrahi (2017)[19]). All these research papers have used Accuracy, Sensitivity, and Specificity as the performance metric to evaluate the models. Table 3 presents the comparison of the results with these research articles and our results. It is found that our proposed approach is outperforming all the existing results in terms of all metrics.

| Research Articles | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| **Xue et al. (2010)** | 88.70 | - | - |
| **Sundarkumar et al. (2015)** | 58.92 | 95.52 | 56.58 |
| **Sundarkumar and Ravi (2015)** | 60.31 | 90.00 | 58.69 |
| **Nian et al. (2016)** | - | 91.00 | 52.00 |
| **Subudhi and Panigrahi (2017)** | 87.02 | 83.21 | 88.45 |
| **Our Proposed Method** | **99.64** | **99.97** | **99.32** |

Table 3: Comparison of proposed method with an existing results.

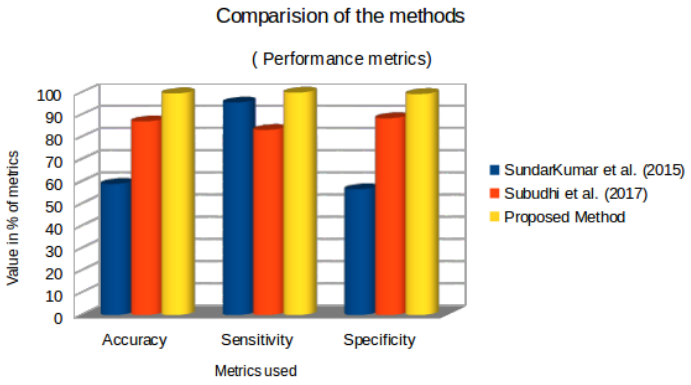


Figure 6: Comparison of our proposed approach

The Figure 6 compares the results of the research articles with our proposed method and finds that our proposed approach gives the highest value in all the three metrics: accuracy, sensitivity and specificity, used.

## 5 Conclusion And Future Work

In this paper, we have proposed a method for building a classifier for detection of fraudulent automobile insurance claims. We have used the MWMOTE, as an over-sampling technique to generate the new synthetic samples and make the dataset symmetric. With the balanced dataset, we have built three different classifiers: Support Vector Machine, Decision Tree and Random Forest. We found that Random Forest was giving the best result among these classifiers. We have also compared our results with the existing results of research articles and found that our results were optimal with respect to all the performance metrics used.

It is found that the oversampling technique MWMOTE wastakingalmost7hourstogeneratethenewsyntheticsamples in this "carclaims.txt" dataset. As the data size becomes bigger, the generation of new synthetic samples takes more time. Hence to reduce the time taken by MWOTE, parallel implementation of MWMOTE on GPU can be done as a part of future work. Building a deep model for automobile insurance fraud detection using deep learning can also be seen as a part of future work.

## References

[1] Sukarna Barua, Md. Monirul Islam, Xin Yao and Kazuyuki, MWMOTE-Majority Weighted Oversampling Technique for Imbalanced Dataset Learning, IEEE Transactions on Knowledge and Data Engineering, Vol.26, No.2, February 2014.

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

[3] Cutting corners, August 2015, Cutting corners to get cheaper motor insurance backfiring on thousands of motorists warns the abi, https://www.insurancefraudbureau.org/mediacentre/news/2015/cutting-corners-to-get-cheaper-motorinsurance-backfiring-on-thousands-of-motorists-warns-the-abi/

[4] ke Nian, Haofan Zhang, Aditya Tayal, Thomas Coleman and Yujing Li, Auto insurance fraud detection using unsupervised spectral ranking for anomaly, The Journal of Finance and Data Science, 2 March,2016,58-75

[5] 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

[6] I. Bose and R. K. Mahapatra, Business data mining-a machine learining perspective, Information and Management, vol. 39, pp.211-225,2001

[7] Salvatore Stolfo, Andreas L. Prodromidis, Shelley Tselepis, Wenke Lee, Dave W. Fan and Philip K. Chan, JAM: Java Agents for Meta-Learning over Distributed Databases, KDD97 Proceedings, 1997

[8] Chan, P.K., Fan, W., Prodromidis, A.L., Stolfo, S.J., Distributed data mining in credit card fraud detection, IEEE Intel. Syst., vol. 14, pp.67-74, 1999

[9] Stolfo, S.J., Fan, D.W., Lee, W., Prodromidis, A.L., Chan, P., Cost-based modelling for fraud and intrusion detection: results from the JAM project.In, Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX), vol. 2, pp. 130-144, 2000

[10] Patrick L. Brockett, Richard A. Derrig, Linda L. Golden, Arnold Levine and Mark Alpert,Fraud Classification Using Principal Component Analysis of RIDITs,The Journal of Risk and Insurance, 2002

[11] Clifton Phua, Damminda Alahakoon, and Vincent Lee,Minority Report in Fraud Detection: Classification of Skewed Data, ACM SIGKDD Explore Newslett, 6(1),pp. 50-59

[12] S. Viaene, S. Viaene and S. Viaene, Auto claim fraud detection using Bayesian learning neural networks,Expert Systems with Applications: An international Journal, vol. 29(3), pp. 653-666, October, 2005

[13] Pathak, J., Vidyarthi, N., Summers, S.L., A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims, Managerial Auditing J., vol. 20(6), pp. 632644, 2005

[14] Bermudez, L., Perez, J., Ayuso, M., Gomez, E., Vazquez, F., A bayesian dichotomous model with asymmetric link for fraud in insurance, Insurance: Math. Econ., vol. 42(2),pp. 779786,2008

[15] Lovro Šubelj, Štefan Furlan and Marko Bajec, An expert system for detecting automobile insurance fraud using social network analysis, Expert Systems with Application, pp. 10391052, 2011

[16] Wei Xu, Shengnan Wang, Dailing Zhang and Bo Yang, Random Rough Subspace Based Neural Network Ensemble for Insurance Fraud Detection, Fourth International Joint Conference on Computational Science and Optimization, IEEE, pp. 1276-1280,2011

[17] G. Ganesh Sundarkumar and Ravi Vadlamani, A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, Engineering Applications of Artificial Intelligence, January, 2015

[18] Ke Nian, Haofan Zhang, Aditya Tayal, Thomas Coleman and Yuying Li, Auto insurance fraud detection using unsupervised spectral ranking for anomaly, The Journal of Finance and Data Science, pp. 58-75, 2016

[19] Ke Nian, Haofan Zhang, Aditya Tayal, Thomas Coleman and Yuying Li, Auto insurance fraud detection using unsupervised spectral ranking for anomaly, The Journal of Finance and Data Science, pp. 58-75, 2016

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

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

[22] P. Hart, The condensed nearest neighbor rule (Corresp.), IEEE Transaction on Information Theory, vol. 12(3), 1968
[23] M.Sternberg and R.G.Reynolds, Using cultural algorithms to support re-engineering o frule-based expert systems in dynamic performance environments: a case study in fraud detection, Evolutionary Computation, IEEE Transactions, vol. 1, pp. 225243, 1997.
[24] Laurikkala, J., Improving identification of difficult small classes by balancing class distribution, Proceedings of the 8th Conference on AI in Medicine. pp. 63–66, 2011
[25] Chawla, N.V., Bower, K.W., Hall, L.O. and Kegelmeyer, W.P., SMOTE:Syntheticminorityover-samplingtechnique,Journalof Artifical Intelligence Research, vol. 16, pp. 321-357, 2002
[26] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008
[27] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, BorderlineSMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, International Conference on Intelligent Computing, pp. 878-887,2005
[28] G. Ganesh Sundarkumar, Vadlamani Ravi and V. Siddeshwar, One-Class Support Vector Machine based undersampling: Application to Churn prediction and Insurance Fraud detection, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2015
[29] G. Ganesh Sundarkumar, Vadlamani Ravi and V. Siddeshwar, One-Class Support Vector Machine based undersampling: Application to Churn prediction and Insurance Fraud detection, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2015
[30] Georgios Douzas1 and Fernando Bacao, Effective data generation for imbalanced learning using Conditional Generative Adversarial Networks, Expert Systems with Applications, vol. 91, pp. 464-471, 2018