Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models

Mohamed Hanafy Kotb[1], Ruixing Ming[2]
School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China[1]
Department of Statistics, Mathematics, and Insurance, Faculty of Commerce, Assuit University, Asyut 71515, Egypt[1]
School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China[2]

***Abstract*—Default in premium payments impacts significantly on the profitability of the insurance company. Therefore, predicting defaults in advance is very important for insurance companies. Predicting in the insurance sector is one of the most beneficial and important study areas in today's world, thanks to technological advancements. But because of the imbalanced datasets in this industry, predicting insurance premium defaulting becomes a difficult task. Moreover, there is no study that applies and compares different SMOTE family approaches to address the issue of imbalanced data. So, this study aims to compare different SMOTE family approaches. Such as Synthetic Minority Oversampling Technique (MOTE), Safe-level SMOTE (SLS), Relocating Safe-level SMOTE (RSLS), Density-based SMOTE (DBSMOTE), Borderline-SMOTE(BLSMOTE), Adaptive Synthetic Sampling (ADSYN), and Adaptive Neighbor Synthetic (ASN), SMOTE-Tomek, and SMOTE-ENN, to solve the problem of unbalanced data. This study applied a variety of machine learning (ML)classifiers to assess the performance of the SMOTE family in addressing the imbalanced problem. These classifiers including Logistic Regression (LR), CART, C4.5, C5.0, Support Vector Machine (SVM), Random Forest (RF), Bagged CART(BC), AdaBoost (ADA), Stochastic Gradient Boosting, (SGB), XGBOOST(XGB), NAÏVE BAYES, (NB), k-Nearest Neighbors (K-NN), and Neural Networks (NN). Additionally, model validation strategies include Random hold-out. The findings obtained using various assessment measures show that ML algorithms do not perform well with imbalanced data, indicating that the problem of imbalanced data must be addressed. On the other hand, using balanced datasets created by SMOTE family techniques improves the performance of classifiers. Moreover, the Friedman test, a statistical significance test, further confirms that the hybrid SMOTE family methods are better than others, especially the SMOTE -TOMEK, which performs better than other resampling approaches. Moreover, among ML algorithms, the SVM model has produced the best results with the SMOTE- TOMEK.**

***Keywords*—Machine learning; classification; insurance; imbalanced data; SMOTE family; statistical analysis**

I. Introduction

In the era of the industrial revolution, all businesses seek digital transformation. One of the key elements of digital transformation is your ability to manage data. Data Science and business analytics is the tool that is being employed on the holy grail of data to extract hidden insights. Since the amount of data is exponentially increasing, therefore the systematic process of data science is gaining popularity in recent times. Like any other industry, 'THE INSURANCE' industry is no exception, and in fact, it is one of the key areas where data science is being practiced at a large scale. Many insurance companies are now employing ML techniques that provide a more systematic way of obtaining a more accurate and representative outcome than the traditional statistic approach.

One of the main challenges with ML approaches in classification is that they are influenced by the data set's unequal class distribution. In other words, when the data is uneven, many ML algorithms may simply disregard the tiny class and assign the majority of the cases to the common class, resulting in high overall model accuracy. Still, the prediction models' efficiency for the tiny class will be drastically diminished. Thus, this study aims to apply a variety of SMOTE family techniques to deal with the imbalanced data problem to improve the performance of ML models in predicting the small class efficiently. In our study, we will develop 117 ML models for predicting insurance premium defaulting {(9 of SMOTE family methods) × (13 of ML models) = 117 model}.

The following is the structure of this paper: Section II presents the previous studies. Section III explains the methodology included data collection, Data Preparation, and imbalanced data problem. Section IV explains model training and parameter optimization. Section V presents the evaluation methods. Section VI shows the results. Section VII shows the results of the statistical tests. Section VIII and IX represent the conclusion and the future work, respectively.

II. Related Work

In the study of [1], they employed several data level methodologies to try to address the unbalanced data issue to predict the occurrence of claims in insurance. The AdaBoost model with oversampling and the hybrid technique produced the highest accurate results. And [2]; they used big insurance data to build eight ML algorithms to predict the occurrence of claims, and they handled the highly imbalanced data using the over-sampler technique. The random forest classifier outperformed the other algorithms. Furthermore, [3] constructed a model for forecasting insurance claims; they generated four classifiers to predict the claims, with the XGBoost model outperforming the others. And [4] predicted

the frequency of vehicle insurance claims using two competing approaches, logistic regression and XGBoost. According to this study, the XGBoost model outperforms logistic regression. Further, the [5] study is to investigate data mining approaches for developing a predictive classifier for vehicle insurance claim prediction. Their studies revealed that neural networks were the best predictor. And [6], this study intends to provide an accurate way for insurance companies to forecast whether or not the customer relationship with the insurance company will be renewed or not. In this paper, random forests were shown to be the top-performing algorithm. And [7], this study starts with data enrichment and works its way up to model development to predict customer churn. And they applied class weights to the prediction model due to the imbalance of the samples. And in [8] the aim of this paper is to compare and contrast the results of different machine-learning techniques for churn prediction; according to the results of this study, the Random Forest and ADA improve outperform all other methods. The study of [9] shows that after using resampling techniques to solve the imbalanced data problem, the efficiency of all ML classifiers in predicting auto insurance fraud is enhanced. Besides, the Stochastic gradient boosting classifier obtained the best result after using the SMOTE-ENN resampling technique among all the other models. And [10] created a new approach for improving the accuracy of fraud prediction. And to solve the unbalanced data problem, they re-balance the data through the method "Resample" of Weka before applying testing and learning. According to this study, Random Forest outperforms all other algorithms in terms of fraud prediction. And [11] predicts fraudulent claims and estimates insurance premium amounts for a range of customers depending on their personal and financial data. The results showed that the Random Forest outperforms the other two algorithms on the Insurance claim dataset. And to deal with the unbalanced data distribution, the research of [12] provides a novel insurance fraud detection technique. The paper is based on constructing insurance fraud detection models based on data partitions derived from under-sampling. The results show that DT outperforms other algorithms.

To accentuate the importance of our study and the gap that we will fill in this study, we summarized a list of recent research that works on classification in the insurance industry by applying the ML models is presented in Table I.

TABLE I. REVIEW OF RESEARCH WORKS IN THE FIELD OF CLASSIFICATION IN THE INSURANCE INDUSTRY

The study	ML models											SMOTE FAMILY									Statistical analysis
	LR	Decision tree (CART or C4.5, or C5.0)	SVM	RF	Bagged CART	ADA	SGB	XGBOOST	NB	k-NN	NN	ADASYN	ANS	BLSMOTE	DBSMOTE	RSLS	SLS	SMOTE	SMOTE-ENN	SMOTE-Tomek	
[1]		CART, C5.0, C4.5		√	√	√	√	√										√			
[2]	√	CART, C5.0, C4.5		√				√	√	√											
[3]		C4.5						√	√		√										
[4]	√							√													
[5]	√	C4.5									√										
[6]	√		√	√		√				√	√										
[7]				√																	
[8]	√	CART	√	√		√	√		√	√	√										
[9]	√	CART, C5.0, C4.5	√	√		√	√	√	√	√	√							√	√		
[10]	√	C4.5	√	√		√			√		√										
[11]		C4.5		√					√												
[12]		√	√								√										
Present study	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

Table I demonstrates that there is an absence of application and detailed comparison of the common SMOTE family approaches for handling unbalanced data in the insurance industry. This research aims to look into the impact of SMOTE family techniques on boosting the performance of machine learning models in the insurance industry. So, in this study, we applied numerous SMOTE family approaches for solving the imbalanced data problem to fill in the gaps in the previous studies. As compared to earlier studies, the following are our study's original advances and key procedures:

- Using feature scaling to standardize different data features.
- Implementing and comparing different SMOTE family techniques, including nine different methods.
- Hold-Out is applied as a prominent cross-validation algorithm to perform the validation process.
- Comparison of the efficiency of SMOTE family techniques using different ML algorithms, including 13 different models.
- Using various evaluation approaches, such as Accuracy, sensitivity, specificity, and AUC, to assess the performance of the developed models.
- Showing how the various SMOTE family strategies affect the performance of classifiers.
- Using the Friedman test to analyze the differences among several SMOTE family approaches and indicating the best method among the others.

III. Methodology

This study compares various SMOTE family approaches to handle the imbalanced data problem to discover the optimal methodology and classifier for forecasting insurance premium defaulting. The following are the methodology steps used to attain the objectives of this paper:

- Data Gathering.
- Data Preparation.
- Implementing SMOTE family techniques to solve the issue of the Imbalanced data.
- applying ML classification algorithms.
- Analyzing the outcomes.

Fig. 1 shows the Flow chart of the proposed work in our study.

A. Data Collection

This research has used datasets from an insurance company of Egypt. between 2014 and 2020 years. This data collection has a number of variables that can influence insurance premium defaulting. This dataset includes information on the 93520 clients with ten various features. There are four categorical variables (area type, Accommodation, Marital status, Default or Not), and six continuous &discrete variables (Age Income, Number of Vehicles owned, number of Late payments, number of premiums that paid, Premium amount, the number of dependents for the insured) with no missing values and columns.

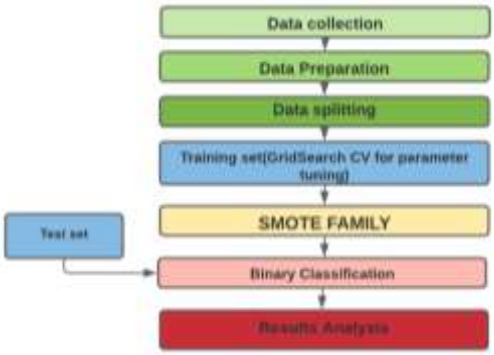

Fig. 1. Working Diagram of Proposed Model.

B. Data Preparation

One of the most crucial stages in ML is data preparation. This procedure turns raw data into an understandable format. This phase will eliminate the errors, which may exist in the dataset, making datasets easier to manage [2]. And the data preprocessing can be summarized into the following two steps.

1) Feature scaling: Feature scaling is a method of normalizing the range of independent variables in a dataset. Most ML algorithms employ the Euclidean distance between two data points, hence without Feature Scaling, the ML algorithms may not perform properly [13]. And in this study, the values range in our dataset is not similar for most variables, so we will apply the Standardization technique as a feature scaling method to rescale the data variables. As a consequence, all of the variables become to have a mean of zero and a standard deviation of one, which is typical of a normal distribution.

The data were scaled using the following algorithm:

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - \mu}{\boldsymbol{\sigma}}$$

Where μ is the mean and $\boldsymbol{\sigma}$ is the standard deviation.

2) One-hot encoding for categorical features: In machine learning, one-hot encoding is the process of converting categorical data into a format that can be fed into ML algorithms. Because most of the ML models works only with the numerical inputs.

C. Imbalanced Data Problem

It's worth noting that most ML algorithms in classification operate best when each class's number of instances is roughly equal. Because the unbalanced data lead to the majority class dominates the minority class. Consequently, algorithms are biased toward the majority class, and their performance become is unreliable [1,14,15]. Our datasets are severely uneven, and the two categories of insurance premium defaulting are not equivalent; in reality, the dataset contains more samples from non-defaulted (90% of the observations)

and defaulted classes (only 10 % of observations). Several techniques have been proposed to address the issue of imbalanced data, SMOTE family is one of the highly effective strategies for resolving the issue of imbalanced data.

SMOTE family: Is a collection of numerous oversampling techniques evolved from SMOTE.

1) Synthetic Minority Oversampling Technique (SMOTE): SMOTE is a statistical strategy that generates new instances to increase the number of minority samples in the dataset. This approach takes feature space samples for each target class and its nearest neighbours, then generates new samples that blend the features from the target case with the features from its neighbours. The new cases are not exact replicas of extant minority cases [16].

2) Adaptive Synthetic Sampling (ADASYN): ADASYN's core concept is to apply a weighted distribution for different minority class instances according to the possibility of learning them. With more artificial instances generated for the minority class instances that are harder to learn than minority class instances that are simpler to learn. Consequently, this technique enhances data distribution learning by eliminating or decreasing the bias brought on by data imbalanced and adaptively pushing the classification decision boundary toward difficult instances [17].

3) Borderline-SMOTE (BLSMOTE): BLSMOTE is a new minority over-sampling technique founded on the SMOTE method that over-samples only the minority examples at the borderline, where the number of majority neighbours of each minority instance is used to split minority instances into three groups: SAFE/DANGER/NOISE. Only the DANGER is employed to generate synthetic instances [18].

4) Density-based SMOTE (DBSMOTE): DBSMOTE, a new over-sampling approach. This method is based on a density-based clustering concept and is intended to oversample a randomly shaped cluster obtained by DBSCAN. DBSMOTE creates synthetic instances by finding the shortest path between each positive instance and a minority-class cluster's pseudo centroid. As a result, the synthetic dataset that results are dense around the core of a group of original positive cases [19].

5) Adaptive Neighbor Synthetic (ANS): The requirement of the number of nearest neighbours as a critical parameter to synthesize instances is one of SMOTE's drawbacks. And The Adaptive Neighbor Synthetic Minority Oversampling Technique (ANS) is a new adaptive technique that tries to avoid this drawback by dynamically adapts the number of neighbours required for oversampling around different minority regions [20].

6) Safe-level SMOTE (SLS): SMOTE synthesizes minority instances at random along a line connecting a minority instance, and it's chosen nearest neighbours while disregarding surrounding majority instances. SLS is a technique that meticulously samples minority instances along the same line with varied weight degrees, which is referred to as the safe level. The safe level is calculated using the minority instances of the nearest neighbours [21].

7) Relocating Safe-level SMOTE (RSLS): SLS creates synthetic minority instances in the vicinity of original instances while avoiding majority instances nearby. This may cause some classifiers to become confused. Furthermore, SLS generates synthetic instances without employing minority outcast instances; thus, some valuable information of the minority class may be lost in the dataset. And by merging two methods, the RSLS tries to address these two flaws in SLS. The first is to check and move these synthetic instances away from any potentially nearby majority instances. The second is using the 1-nearest neighbour strategy to deal with minority outcasts [22].

8) HYBRID techniques: smote family that are considered as over-sampling methods have their own set of benefits and drawbacks. Combining the Over-sampling methods with the under-sampling can help reap the benefits of both.

a) SMOTE-ENN: The SMOTE-ENN technique is one of the most well-known techniques for improving outcomes by combining the SMOTE that represent an over-sampling technique with the Edited Nearest Neighbors (ENN) that represent an under-sampling technique [23].

b) SMOTE-Tomek: The SMOTE-Tomek technique combines the SMOTE that represents an over-sampling technique with the Tomek that represents an under-sampling technique to improve outcomes [23].

IV. Model Training with Parameter Optimization

A. Model Validation

By using the cross-validation technique, the data were divided into training and testing subsets. Cross-validation of input data is used to prevent machine learning models from overfitting and underfitting. This study used the Random holdout as a popular cross-validation procedure.

You can see a scheme of holdout CV in Fig. 2

Fig. 2. Holdout CV.

- The data is randomly split into a training and test set.
- A model is trained using only the training set.
- Predictions are made on the test set.
- The predictions are compared to the true values.

B. Overfitting and Underfitting

Machine model's training and validation scores will be recorded at lower levels in the case of Underfitting. In comparison, overfitting is defined as a pattern of high training scores combined with low validation results. Model parameters must be optimized to avoid overfitting and underfitting circumstances. The grid search technique, which is a popular tuning tool, was used to optimize the parameters of the models. Table II shows the best values for model parameters.

TABLE II. MACHINE LEARNING MODELS WITH THEIR SPECIFIC PARAMETER'S SETTINGS

K-NN	K=23	SVM	C =0.8
CART	cp = 0.006329114.	RF	mtry = 2
C4.5	C = 0.01. M = 5.	NN	size = 1. decay = 0.1
LR	no tuning parameters.	ADA	nIter = 150. method = Real adaboost.
NB	laplace = 0. usekernel = TRUE. adjust = 0.4.	C5.0	trials = 90. model = tree. winnow = FALSE
XGB	nrounds = 50. max_depth = 2. eta = 0.3. gamma = 0. colsample_bytree = 0.8. min_child_weight = 1. subsample = 1.	SGB	n. trees = 50. interaction. depth = 1. shrinkage = 0.1. n. minobsinnode = 10.
BC	no tuning parameters.		

V. EVALUATION METHODS

Methods of evaluation are critical in comparing and selecting the best model [1].

TABLE III. EVALUATION METHODS

Accuracy	Referred to the overall correctly prediction	$\frac{(TP + TN)}{(TP + FP + TN + FN)}$
Sensitivity	Referred to the correct rate of predicting the default class.	$\frac{TP}{(TP + FN)}$
Specificity	Referred to the correct rate of predicting the non-default class	$\frac{TN}{(FP + TN).}$

The evaluation methods employed in this study are shown in Table III. Where TP is the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives.

Where:

1) TP: is the aggregate number of clients who accurately attributed to default class.

2) FP: is the aggregate number of clients who inaccurately attributed to the default class.

3) TN: is the aggregate number of clients who accurately attributed to non-default class.

4) FN: is the aggregate number of clients who inaccurately attributed to the non-default class.

Besides the evaluation methods in Table III, we also used the AUC, AUC is a universal quality metric for models. AUC of 1 indicates a perfect model, whereas an AUC of 0.5 indicates a random model.

Analyzing and comparing the performance of the classifiers is an important procedure. Although evaluation measures are straightforward to employ, the results obtaining from the evaluation measures may be misleading. As a result, determining the optimal model or technique according to their abilities is a difficult task. This problem will be solved using statistical significance tests [24]. A common statistical test method for determining the differences between two or more related sample means is called the ANOVA test. The ANOVA's null hypothesis is that all resampling procedures are equivalent, and the stated discrepancies are just coincidental [25]. There are three assumptions that must take into account before we applied the ANOVA test.

1) All samples must follow the normal distribution.

2) The sample cases should be independent of one another.

3) There should be roughly equal variance among the methods (SMOTE family methods).

The Anderson–Darling normality test [25] is used in this study to determine whether data is normal or not. The null hypothesis of this Anderson–Darling normality test is that the data follow a normal distribution. And we will accept this null hypothesis if the p-value of the test is more than 0.05; otherwise, we will reject the null hypothesis if the p-value $\leq$ 0.05.

If one of the ANOVA's assumptions be broken, the Friedman test [26] will be used instead of the ANOVA test to investigate differences among the methods. The Friedman test's null hypothesis is that all SMOTE family methods perform the same. And we will accept the null hypothesis if the p-value of the test is more than 0.05; otherwise, we will reject the null hypothesis if the p-value $\leq$ 0.05. And rejecting the null hypothesis means that at least one of the SMOTE family strategies perform differently from others. For each SMOTE family approach, the accuracy, sensitivity, specificity, and AUC values are used to compare the ability of the different resampling techniques to tackle the problem of unbalanced data.

The Freidman test ranks each classifier's data for each SMOTE family technique, then examines the ranks values [27].

As a result, for each SMOTE family technique, the Friedman test generates a sum of ranks, which aids in determining which SMOTE family method is the most effective among the others.

VI. RESULTS

The performance of the various ML classifiers on the unbalanced dataset and also on the balanced data that was generated by the SMOTE family methods is shown in Table IV. Various assessment measure methods, including accuracy, sensitivity, specificity, and AUC, are utilized to gain a better knowledge of the models' performance.

Table IV shows the accuracy, sensitivity, specificity, and AUC of each ML strategy on balanced and imbalanced datasets created by the SMOTE family. The most important outcomes are from Table IV; there is a substantial discrepancy between specificity and sensitivity with the unbalanced data.

TABLE IV. PERFORMANCE OF THE CLASSIFIERS

ML	Evaluation	unbalanced	ADASYN	ANS	BLSMOTE	DBSMOTE	RSLS	SLS	SMOTE	SOMTE – TOMEK	SMOTE - ENN
K-NN	Accuracy	0.9105	0.8023	0.8092	0.8583	0.8363	0.8737	0.8725	0.8035	0.6654	0.7283
	sensitivity	0.11911	0.5198	0.5128	0.4013	0.4118	0.3247	0.3142	0.5024	0.7408	0.6607
	specificity	0.96489	0.8243	0.832	0.8914	0.8674	0.9127	0.912	0.8267	0.6625	0.7345
	AUC	0.542	0.67205	0.6724	0.64635	0.6396	0.6187	0.6131	0.66455	0.70165	0.6976
LR	Accuracy	0.9062	0.7791	0.781	0.8074	0.7849	0.7849	0.8787	0.7849	0.7072	0.7767
	sensitivity	0.1727	0.7568	0.769	0.7201	0.7201	0.7201	0.4754	0.7201	0.8349	0.7687
	specificity	0.9638	0.7852	0.7863	0.8182	0.7941	0.7941	0.9124	0.7941	0.701	0.7792
	AUC	0.56825	0.771	0.77765	0.76915	0.7571	0.7571	0.6939	0.7571	0.76795	0.77395
SVM	Accuracy	0.9025	0.776	0.7676	0.7981	0.7733	0.8913	0.8899	0.7684	0.7082	0.7707
	sensitivity	0.032	0.7659	0.7751	0.7262	0.7323	0.4968	0.4418	0.7843	0.8384	0.7722
	specificity	0.97	0.7813	0.7715	0.8078	0.7808	0.9288	0.9268	0.7018	0.7717	0.7725
	AUC	0.501	0.7736	0.7733	0.767	0.75655	0.7128	0.6843	0.74305	0.80505	0.77235
NB	Accuracy	0.9018	0.8539	0.8591	0.856	0.8213	0.8659	0.8667	0.8566	0.8741	0.8826
	sensitivity	0.032	0.5366	0.5305	0.5611	0.5886	0.4357	0.4968	0.5488	0.5213	0.4969
	specificity	0.9693	0.8814	0.8874	0.8818	0.8426	0.8971	0.898	0.8834	0.8985	0.9091
	AUC	0.50065	0.709	0.70895	0.72145	0.7156	0.6664	0.6974	0.7161	0.7099	0.703
C5.0	Accuracy	0.9028	0.898	0.8974	0.8986	0.8955	0.8994	0.8967	0.9003	0.7254	0.8074
	sensitivity	0.2031	0.2031	0.2237	0.2237	0.1913	0.2326	0.2355	0.2178	0.7478	0.6711
	specificity	0.9606	0.9554	0.9532	0.9545	0.9537	0.9548	0.9516	0.9568	0.7258	0.818
	AUC	0.58185	0.57925	0.58845	0.5891	0.5725	0.5937	0.59355	0.5873	0.7368	0.74455
C4.5	Accuracy	0.8988	0.8926	0.8919	0.8907	0.9165	0.9233	0.9247	0.9168	0.7437	0.8074
	sensitivity	0.1264	0.2267	0.2208	0.2031	0.3653	0.5589	0.5425	0.387	0.7269	0.7025
	specificity	0.9622	0.9478	0.9476	0.9476	0.9583	0.9524	0.955	0.9572	0.7467	0.816
	AUC	0.5443	0.58725	0.5842	0.57535	0.6618	0.75565	0.74875	0.6721	0.7368	0.75925
CART	Accuracy	0.898	0.8782	0.8834	0.8811	0.8851	0.8824	0.8795	0.8786	0.7688	0.8271
	sensitivity	0.09395	0.3978	0.3919	0.4243	0.386	0.4007	0.4361	0.4361	0.7617	0.6572
	specificity	0.96373	0.9194	0.9254	0.9205	0.9277	0.9236	0.9178	0.9169	0.7712	0.8398
	AUC	0.52884	0.6586	0.65865	0.6724	0.65685	0.66215	0.67695	0.6765	0.76645	0.7485
BC	Accuracy	0.9057	0.9036	0.9057	0.9061	0.9048	0.9025	0.903	0.904	0.7429	0.7911
	sensitivity	0.1783	0.1992	0.2445	0.2236	0.2062	0.2167	0.2202	0.2202	0.7269	0.6676
	specificity	0.956	0.9524	0.9518	0.9536	0.9533	0.9502	0.9504	0.9516	0.7458	0.8009
	AUC	0.56715	0.5758	0.59815	0.5886	0.57975	0.58345	0.5853	0.5859	0.73635	0.73425
XGB	Accuracy	0.9103	0.9088	0.9107	0.9076	0.9092	0.8925	0.8956	0.9101	0.7527	0.8158
	sensitivity	0.1574	0.147	0.1783	0.1679	0.1749	0.2898	0.2898	0.1714	0.7164	0.6363
	specificity	0.9622	0.9613	0.9613	0.9587	0.96	0.9349	0.9382	0.9611	0.7569	0.8291
	AUC	0.5598	0.55415	0.5698	0.5633	0.56745	0.61235	0.614	0.56625	0.73665	0.7327
ADA	Accuracy	0.9105	0.9089	0.90975	0.9063	0.9091	0.89	0.8909	0.90935	0.7458	0.80745
	sensitivity	0.04245	0.17485	0.19575	0.2045	0.19755	0.34035	0.3351	0.1923	0.7443	0.6694
	specificity	0.96978	0.95965	0.9592	0.955	0.95845	0.929	0.93035	0.959	0.7478	0.81815
	AUC	0.506115	0.56725	0.577475	0.57975	0.578	0.634675	0.632725	0.57565	0.74605	0.7437
SGB	Accuracy	0.9117	0.909	0.9088	0.905	0.909	0.8875	0.8862	0.9086	0.7389	0.7991
	sensitivity	0.1505	0.2027	0.2132	0.2411	0.2202	0.3909	0.3804	0.2132	0.7722	0.7025
	specificity	0.9642	0.958	0.9571	0.9513	0.9569	0.9231	0.9225	0.9569	0.7387	0.8072
	AUC	0.55735	0.58035	0.58515	0.5962	0.58855	0.657	0.65145	0.58505	0.75545	0.75485
RF	Accuracy	0.9107	0.9071	0.9078	0.9096	0.9101	0.9096	0.9071	0.9082	0.751	0.8216
	sensitivity	0.0773	0.1992	0.2341	0.1435	0.11911	0.1505	0.1365	0.2132	0.7478	0.6328
	specificity	0.9678	0.9562	0.9547	0.9624	0.96445	0.962	0.9602	0.9564	0.7532	0.8356
	AUC	0.52255	0.5777	0.5944	0.55295	0.54178	0.55625	0.54835	0.5848	0.7505	0.7342
NN	Accuracy	0.9109	0.6809	0.7676	0.7653	0.7488	0.847	0.8029	0.7728	0.7153	0.7404
	sensitivity	0.1714	0.8334	0.7045	0.6871	0.6278	0.5268	0.6278	0.6836	0.8349	0.7966
	specificity	0.962	0.6752	0.7756	0.7743	0.7605	0.8714	0.818	0.7825	0.7096	0.7387
	AUC	0.5667	0.7543	0.74005	0.7307	0.69415	0.6991	0.7229	0.73305	0.77225	0.76765

From Table IV we can see that in the column of imbalanced Dataset, all of the accuracy results are greater than 90%, all sensitivity values are less than 18 % and all specificity results are greater than 96 %, indicating that all classifiers are biased toward the majority class. So, the problem must be addressed because it led to inaccurate results. And, after using various SMOTE family techniques to solve the unbalanced problem, we can see a significant improvement in the ML systems' ability to forecast the minority class. For example, while utilizing imbalanced data, the SVM got a sensitivity of 3.2 %, but the result increased to 83.84% with the SOMTE -TOMEK technique.

VII. Results of Statistical Tests

The ML algorithms perform differently with the different balanced data created by various SMOTE family techniques. As a result, finding the appropriate SMOTE family approach to get the greatest results from ML algorithms is quite difficult. Thus, we will use a Statistical significance test that will help us in this difficult task of deciding on the optimum SMOTE family technique. And before doing the ANOVA test, it's important to check the normality assumption.

TABLE V. The Results of the Anderson-Darling Normality

Accuracy	A = 6.013,	p-value = 7.099e-15
Sensitivity	A = 3.977,	p-value = 5.834e-10
Specificity	A = 6.3676,	p-value = 1.005e-15
AUC	A = 6.013,	p-value = 7.099e-15

Table V shows the normality test results according to the Anderson-Darling normality test on the accuracy, sensitivity, specificity, and AUC. The p-value is less than 0.05; thus, the null hypothesis is rejected, and the ANOVA test cannot be employed.

Because one of ANOVA's assumptions related to the normal distribution is broken, we will use the Friedman test to compare the resampling strategies in both datasets instead of the ANOVA test. The Friedman test results are shown in Table VI.

TABLE VI. The Friedman Test Results

Accuracy	chi-squared= 40.345	df=8	p-value = 2.763e-06
sensitivity	chi-squared = 35.235	df=8	p-value = 2.423e-05
specificity	chi-squared = 43.959	df=8	p-value = 5.793e-07
AUC	chi-squared = 42.201	df= 8	p-value = 1.242e-06

Table VI shows that the p-value of the Friedman test for Accuracy, Sensitivity, Specificity, and AUC is lower than the (0.05). As a result, we will reject the null hypothesis, and the following conclusion can be drawn at least one of the SMOTE family techniques performs differently from the other methods.

Table VII shows the rank, sum of ranks, and median determined from the Friedman test for Accuracy, Sensitivity, Specificity, and AUC. And Table VII confirm the following results:

1) For the accuracy, the RSLS technique could be more effective than the other techniques

2) For the sensitivity, the DBSMOTE technique could be more effective than the other techniques

3) For the Specificity and the AUC, the SMOTE_TOMEK technique could be more effective than the other techniques.

TABLE VII. Additional Information from Friedman Test Results

	RANK	SMOTE FAMILY	SUM OF RANKS	MEDIAN
Accuracy	1	RSLS	82.5	0.89
	2	DBSMOTE	78.5	0.8955
	3	SMOTE	78	0.9003
	4	BLSMOTE	77.5	0.8907
	5	ANS	77	0.8919
	6	SLS	76.5	0.8899
	7	ADASYN	58	0.8926
	8	SMOTE_ENN	36	0.8074
	9	SMOTE_TOMEK	21	0.7429
		Overall		0.8786
Sensitivity	1	SMOTE_TOMEK	112	0.7478
	2	SMOTE_ENN	95	0.6694
	3	ANS	65	0.2445
	4	SMOTE	60	0.387
	5.5	BLSMOTE	54	0.2411
	5.5	RSLS	54	0.3909
	7	SLS	52	0.3804
	8	ADASYN	49	0.2267
	9	DBSMOTE	44	0.3653
		Overall		0.4361
Specificity	1	DBSMOTE	80.5	0.9533
	2	RSLS	79	0.9288
	3	BLSMOTE	77.5	0.9476
	4	SLS	75	0.9268
	5	SMOTE	72.5	0.9516
	6	ANS	71	0.9476
	7	ADASYN	70.5	0.9478
	8	SMOTE_ENN	36	0.816
	9	SMOTE_TOMEK	23	0.7467
		Overall		0.9127
AUC	1	SMOTE_TOMEK	106	0.74605
	2	SMOTE_ENN	99	0.74455
	3	ANS	67	0.59815
	4	BLSMOTE	60	0.5962
	5.5	RSLS	56	0.657
	5.5	SMOTE	56	0.66455
	7	SLS	53	0.65145
	8	ADASYN	49	0.58725
	9	DBSMOTE	39	0.6396
		Overall		0.6724

Fig. 3, 4, 6 and 5 shows the SMOTE family methods' boxplot based on the accuracy, specificity, sensitivity, and AUC, respectively, where data refers to the original data.

To summarize, in this study, we aim to solve the imbalanced problem with SMOTE family methods; the assessment measures as to the accuracy, sensitivity, specificity, and AUC are utilized to compare models more compactly. Accuracy can be a useful measure if data has the same number of samples per class. However, with an imbalanced set of samples, accuracy is not helpful at all because the model predicts the value of the majority classes for all predictions. So, when it comes to selecting the best models, AUC will take precedence. From Fig. 3, 4, 5 and 6, we can note that ML models achieve the highest accuracy and the highest specificity with the original data. On the other hand, ML models achieve the lowest results for the sensitivity and AUC measures; this refers to ML algorithms do not give accurate results using imbalanced datasets, and they cannot predict all the target classes. Therefore, solving the imbalanced data problem is notably necessary. And by using the balanced dataset after applied SMOTE family, the sensitivity and accuracy-test results are not significantly improved. And it is logical because, on the balanced data, most ML classifiers will consider all classes, which will lead to lower sensitivity and accuracy results than the imbalanced data that considers only one class and ignore the other class. Moreover, the specificity and AUC results using the balanced dataset are significantly improved, especially with the hybrid SMOTE methods.

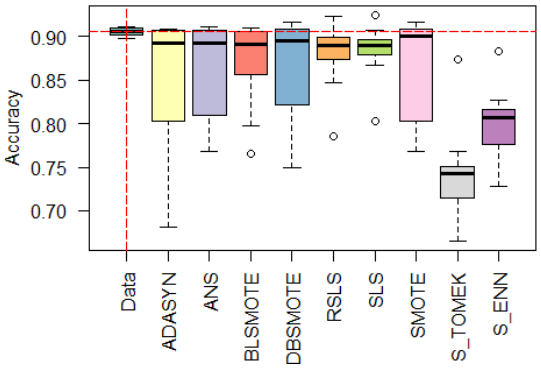

Fig. 3. The Boxplot of the Original Data and SMOTFAMILY based on the Accuracy.

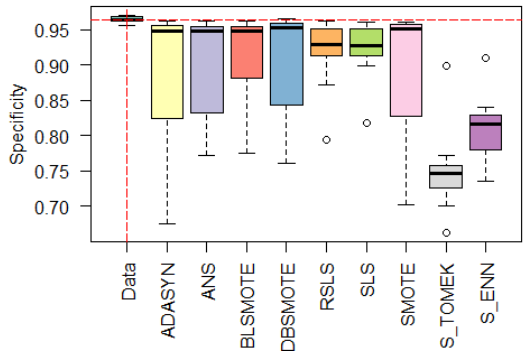

Fig. 4. The Boxplot of the Original Data and SMOTE Family Methods based on the Specificity.

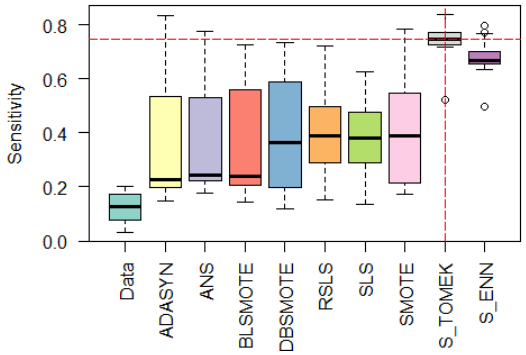

Fig. 5. The Boxplot of the Original Data and SMOTE Family Methods based on the Sensitivity.

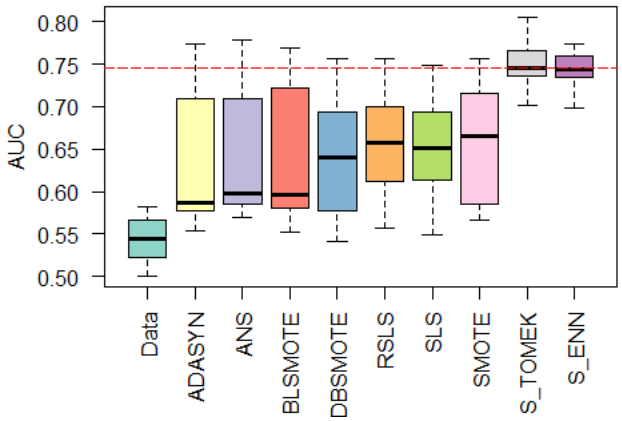

Fig. 6. The Boxplot of the Original Data and SMOTE Family Methods based on the AUC.

Finally, based on the AUC comparison of ML models, the performance of the SVM classifier with the SMOTE-TOMEK method was 80.5%, which was the highest compared with all models.

VIII. Conclusion

The findings show that, algorithms are unable to make accurate predictions with unbalanced data. In contrast, the results demonstrate that algorithms performance has improved when using the various balanced data obtained by different SMOTE family techniques. The findings of the validation approach show that classifiers perform differently on the different balanced data, making it difficult to choose the appropriate resampling technique. The Friedman test was used to determine the optimal resampling approach. According to the AUC, the results of this test show that the hybrid resampling methods are better than others, and especially the SMOTE-TOMEK performs better than alternative resampling approaches. Moreover, among ML algorithms, the SVM model has produced the best results with the SMOTE - TOMEK. According to the results of this paper, we recommend using hybrid resampling strategies to solve the unbalanced data problem as both SMOTE- TOMEK and SMOTE-ENN provided the best performance.

IX. Future Work

The study can be broadened to incorporate hybrid and deep learning algorithms. Other performance indicators might be used to assess performance. The algorithm's timing measures could also be a useful indicator of algorithms performance. Algorithms could also be evaluated with different datasets from various sectors that suffer from the problem of unbalanced data to prove the efficiency of the hybrid resampling strategies to solve the imbalanced data problem.

References

[1] Mohamed Hanafy and Ruixing Ming, "Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches" International Journal of Advanced Computer Science and Applications (IJACSA), 12(6), 2021. http://dx.doi.org/10.14569/IJACSA.2021. 0120656.

[2] Hanafy, Mohamed, and Ruixing Ming. "Machine learning approaches for auto insurance big data." *Risks* 9.2 (2021): 42.

[3] Abdelhadi, Shady, Khaled Elbahnasy, and Mohamed Abdelsalam. "A proposed model to predict auto insurance claims using machine learning techniques." *Journal of Theoretical and Applied Information Technology* 98.22 (2020).

[4] Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression." *Risks* 7.2 (2019): 70.

[5] Weerasinghe, K. P. M. L. P., and M. C. Wijegunasekara. "A comparative study of data mining algorithms in the prediction of auto insurance claims." *European International Journal of Science and Technology* 5.1 (2016): 47-54.

[6] Stucki, Oskar. "Predicting the customer churn with machine learning methods: case: private insurance customer data." (2019). Master's dissertation, LUT University, Lappeenranta, Finland.

[7] Mau, Stefan, Irena Pletikosa, and Joël Wagner. "Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannelenvironments." *International Journal of Bank Marketing* (2018).

[8] Sabbeh, Sahar F. "Machine-learning techniques for customer retention: A comparative study." *International Journal of Advanced Computer Science and Applications* 9.2 (2018).

[9] Hanafy, Mohamed, and Ruixing Ming. "Using Machine Learning Models to Compare Various Resampling Methods in Predicting Insurance Fraud." *Journal of Theoretical and Applied Information Technology* 99.12 (2021).

[10] Itri, Bouzgarne, et al. "Performance comparative study of machine learning algorithms for automobile insurance fraud detection." *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*. IEEE, 2019.

[11] Kowshalya, G., and M. Nandhini. "Predicting fraudulent claims in automobile insurance." *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018.

[12] Hassan, Amira Kamil Ibrahim, and Ajith Abraham. "Modeling insurance fraud detection using imbalanced data classification." *Advances in nature and biologically inspired computing*. Springer, Cham, 2016. 117-127.

[13] S. Aksoy and R. M. Haralick, ''Feature normalization and likelihood-based similarity measures for image retrieval,'' Pattern Recognit. Lett., vol. 22, no. 5, pp. 563–582, Apr. 2001.

[14] H. Byeon, Development of a physical impairment prediction model for Korean elderly people using synthetic minority over-sampling technique and XGBoost. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 36-41, 2021.

[15] H. Byeon, Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 74-79, 2021.

[16] Siriseriwan, Wacharasak. "SMOTEfamily: a collection of oversampling techniques for class imbalance problem based on SMOTE (2018)." *URL http://cran. r-project. org/package= SMOTEfamily*.

[17] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", *Proc. IEEE Int. Joint Conf. Neural Netw. IEEE World Congr. Comput. Intell.*, pp. 1322-1328, Mar. 2008.

[18] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." *International conference on intelligent computing*. Springer, Berlin, Heidelberg, 2005.

[19] Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "DBSMOTE: density-based synthetic minority over-sampling technique." *Applied Intelligence* 36.3 (2012): 664-684.

[20] Siriseriwan, Wacharasak, and Krung Sinapiromsaran. "Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling." *Songklanakarin J. Sci. Technol* 39.5 (2017): 565-576.

[21] Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." *Pacific-Asia conference on knowledge discovery and data mining*. Springer, Berlin, Heidelberg, 2009.

[22] Siriseriwan, Wacharasak, and Krung Sinapiromsaran. "The effective redistribution for imbalance dataset: relocating safe-level SMOTE with minority outcast handling." *Chiang Mai Journal of Science* 43.1 (2016): 234-246.

[23] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." *ACM SIGKDD explorations newsletter* 6.1 (2004): 20-29.

[24] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." *The Journal of Machine Learning Research* 7 (2006): 1-30.

[25] R. A. Fisher, Statistical Methods and Scientific Inference. Oxford, U.K.: Hafner Publishing Co, 1956.

[26] Friedman, Milton. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." *Journal of the american statistical association* 32.200 (1937): 675-701.

[27] Friedman, Milton. "A comparison of alternative tests of significance for the problem of m rankings." *The Annals of Mathematical Statistics* 11.1 (1940): 86-92.