Motor Insurance Claim Status Prediction using Machine Learning Techniques

Endalew Alamir[1], Teklu Urgessa[2], Ashebir Hunegnaw[3], Tiruveedula Gopikrishna[4]
Department of Management Information Systems, Mettu University, Mettu, Ethiopia[1, 3]
Department of Computer Science and Engineering, Adama Science and Technology University, Adama, Ethiopia[2, 4]

***Abstract*—The insurance claim is a basic problem in insurance companies. Insurance insurers always have a challenge to the growing of insurance claim loss. Because there is the occurrence of claim fraud and the volume of claim data increases in the insurance companies. As a result, it is difficult to classify the insured claim status during the claim review process. Therefore, the aims of the study was to build a machine learning model that classifies and make motor insurance claim status prediction in machine learning approach. To achieve this study Missing value ratio, Z- Score, encoding techniques and entropy were used as data set preparation techniques. The final preprocessed data sets split using K- Fold cross validation techniques into training and testing sets. Finally the prediction model was built using Random Forest (RF) and Multi Class – Support Vector Machine (SVM).The performance of the models, RF and Multi –Class SVM classifiers were evaluated using Accuracy, Precision, Recall, and F- measure. The prediction accuracy of the model is capable of predicting the motor insurance claim status with 98.36% and 98.17% by RF and SVM classifiers respectively. As a result, RF classifier is slightly better than Multi-Class Support vector machines. Developing and implementing hybrid model to benefit from the advantages of different algorithms having graphical user interface to apply the solution to real world problem of the insurance company is a pressing future work.**

Keywords*—Motor insurance claim; machine learning; classification; Random Forest (RF); Support Vector Machine (SVM); supervised learning

I. INTRODUCTION

Insurance company is fast growing, industry [1] [2]. It has great role in assuring economic wellbeing of a country, and Insurance claims in insurance companies are costly problems [3]. Insurance providers always make a great effort, with the growing of insurance claim cost or claim loss because of insurance claim fraud [4]. Insurance companies have business problems, such as risk assessment, classification of policy holders and resource allocation, insurance claim classification and prediction in the insurance claim handling process [3]. This insurance business problems were not solved using traditional analytical approaches, including regression, linear programming [5].

Nowadays an insurance corporation has been struggled (stressed) to get best methods that handle transactional data and, risk management data for years [6]. But there is a recent emphasis to use different sources, of data which extends beyond traditional data sources, often known as big data. This big data has created to change data management across the insurance industry [7] [8]. Data variety and data volume push the traditional data management (Relational Database Management System (RDBMS) technologies and software tools because of their restrictions [7] [9].

As the computing technology has been technologically advanced enormously [5], machine learning approach is used to solve insurance business problems like insurance risk, claim loss, to understand and analysis huge amount of data [10] [11]. Companies have huge amounts of data, in the insurance database, which could not be understandable and interpretable by humans like Ethiopian Insurance companies specifically Awash motor insurance claim data.

Therefore, handling and processing large amount of insurance claim data requires computational tools. Machine learning approaches are essential to process the data and, extract the vital insurance claim information for decision making process [5] [12].

For these problems, supervised machine learning techniques, particularly classification algorithms are used as the computational processes for the data set that stored in the insurance database. Machine learning classifiers are used to classify different types or classes of data from a dataset to predict what will happen in the future from the past data set [5] [11].

Machine learning approach in big data is helping to connect machine with huge databases making them to learn new things by its own. Analysis of big data using machine learning approach helps the insurance industry to predict future trends in the competitive market. Big data initially emerged as a term in order to describe data sets whose amount or size is beyond the capability of traditional databases, to capture, store, analyze, manage, and too complex to analyze by traditional data processing techniques and database management tools [9] [13]. Big data is not only about the size, finding insights from complex, heterogeneous, and complex, noisy and voluminous data [11]. Big data categorized as structured data, unstructured data and semi structured data. Structured data is accessed, stored and processed in the fixed format. The type of data in this study is structured data. Because the motor insurance claims data have stored in fixed format, which is store in fixed relational database format. The main objective of the study was to build machine learning model that classifies and make motor insurance claim status prediction in machine learning techniques.

Finally the proposed motor insurance claim status prediction model was addressed the following research questions.

- Can we build more accurate machine learning model that classify motor insurance claim data and make claim status prediction for the insurance company?
- Which techniques needed to prepare the data sets to be able to apply model building techniques?
- What are the better classification techniques that would use for claim classification and how we evaluate the performance of the built machine learning model?

II. Related Works

This section described the existing related work that has been done before by other researchers .This section includes methods and techniques, implementation tools, aims of study and findings of the research as follows in the following Table I.

TABLE I. Related Works of the Study

Objective of Study	***Methods and Techniques***	***Data and place***	***Findings***
Build Predictive Model for Auto Insurance Claims prediction [18]	CART, Entropy Gini index Decision Tree	1,528 Ghana insurance data Vehicle age and customers age are most predictor variable	Policy holders whose age is 18 to 48 have max claim Vehicle age 0 to 8 years have max claim
Support vector machines to classify policy holders satisfactory in automobile insurance[11][17]	Machine learning algorithm, SVM kernel trick, RBF Parameter 0.05	13,635 Indonesia automobile insurance policies,40% data to train,60% data to test	Classification of Customer satisfaction had claim or not. Reliable SVM model to predict, claim ,84.08% of accuracy
An Ensemble Random Forest Algorithm for Insurance Big Data Analysis[6] [11]	Apache Hadoop, Map reduce Apache spark Ensemble RF SVM,LR Precision , G-mean F-measure ,Information gain	500,000, customers data from China insurance	Ensemble RF Algorithm is better than SVM, and logistic regression for insurance product and policy holder analysis Application of ensemble RF with spark for insurance big data analysis
Data mining classification model to Predict the customer's claims in auto insurance company[2]	Logistics regression, Artificial Neural network, Decision Tree C4.5,Accuracy ,precision, recall	80% sample data as training and 20% sample data as testing	The insurance claims classified as low, high, fair. Neural network Has best prediction accuracy of 61.7% to classify claims
Predict the customer's choice of car insurance policies using random forest[12]	Data mining classifications algorithms include Decision Tree, K-Nearest Neighbors Naïve Bayes, Neural Networks and, and Support vector machine algorithms, weka	665,250 records of insurance policies from Allstate insurance company. 665,250 as train set and 198,857as test set.	split the data in to seven categories in order to predict the customer's car insurance policy The performance of the Random Forest model was 97.9%.

III. Materials and Methods

A. Development Tools

Anaconda Navigator and python programing language was used for this research. Anaconda Navigator tool, Jupiter notebook, scikit – learn (sklearn) frame work, and python programing language was used to implement the proposed model. Descriptive statistics summary and graphics data analysis techniques were used. Descriptive statistics used for motor insurance claim data analysis using count, mean, standard deviation, quartiles (25%, 50%, and 75%), min and max. Graphics techniques were used for visualization of the data distribution, using graphical representation like density plot, histograms, table and bar graph.

B. Data Collection

The sources of data for this research were secondary and primary data sources. Secondary data was collected from the existing centralized insurance database of Awash insurance company main office, which is found at Addis Ababa. The relevant secondary motor insurance claim data were collected from the standard experts of Awash insurance company. In addition to, this the researcher used interview methods in order to understand the insurance domain knowledge and motor insurance claim data with insurance experts of the company.

C. Dataset Description

The amount of the dataset used for this research consists of a sample of 65,535 records or instances of AIC motor insurance claim data. The data set contains a total of eleven attributes of motor insurance claim data. This data has excel data format. The column shows the attributes and the row shows the records (instances). The motor insurance dataset have five target classes of insurance policy holders claim status which are close, notification, pending, re-open and settled. The other ten features (attributes) are policy number, name of insured, claim numbers, claim date, estimated loss, claim paid(gross), net of recoveries, total claims expense paid, change in outstanding and claim incurred. The period of the sample motor insurance claim dataset was covered from 2014 up to 2017. This range takes as a base line of the study, because the AIC started to use system for register insurance claim data at the end of 2013. After a year the system starts to store well organized data in the insurance database.

D. Data Preparation Techniques

Data processing techniques were used for data set preparation. Data preprocessing techniques include: data cleaning, data integration, data normalization or data transformations, and encode as shown in Fig. 2. Data cleaning was used to remove noisy data, irrelevant data, which are 47 non-relevant columns from the data set, and reduce the dimension of the dataset from 58 columns to 11 columns by using dimensional reduction techniques specifically missing value ratio. z - Score was used for data normalization, because it normalizes each feature to have mean of zero and variance of one. It also tells as how many standard deviations each feature far away from the mean and it can normalize the data when the actual min and max value is not known. The formula of z - score described below as equation 1.

$$\sum_{n=1}^{n}\left(\frac{x}{\mathrm{n}}\right)$$

$$\sigma^2 = \sum_{n=1}^{n}\left(\frac{(X - \mathrm{X}')2}{n-1}\right)$$

$$\mathrm{z} = \frac{(Xi - \mathrm{X}')}{\sigma} \quad (1)$$

Where X' is mean, sigma is standard deviations, and Z is Z – Score.

To encode categorical data one – hot encoding (OHE) technique was used to convert claim status categorical data to numeric or binary, because there is no natural ordinal relationship between claim status (closed, notification, pending, re-open, and settled).

Policy Number, Name of Insured ,and Claim Number contains string values as an instances or records, this three features have quantized to numeric data values to make the data understandably by RF, and SVM machine learning algorithm. The other features have numeric and float values, namely Claim paid (gross paid=A), Net of Recoveries=B, Net of Recoveries (A-B), Change in Outstanding. These values have a large difference between the max and min values for each feature. Because of this Z - score data normalization technique was applied to transform or scale down the data set. The last features, which is claim status is encoded by using a label encoder because it is a nominal categorical data. Where the claim status 0, 1, 2, 3, and 4 referrers to Closed, Notification, Pending, Re-open, and settled, respectively.

Attribute evaluation techniques or variable importance measure was used to identify the most relevant attribute or features from the whole attributes during classification process for model construction. For variable importance measure information gain or entropy and domain experts was used.

$$\text{Gain (D, A)=Entropy(D)} - \sum_{j=1}^{v} \frac{|Dj|}{|D|} entropy(Dj) \quad (2)$$

Where D is the data partition, A is attribute, V is partition the instances to D1, D2….. Dj but the entropy can be calculated as follows below, and attribute Aj that have maximum information gain is used as important features .

$$H = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p(xi)} \log_2 \mathrm{p(xi)} \quad (3)$$

Where (pxi) is the probability of selected class and n is number of the data set class and H is entropy. The following Fig. 1 shows the relative importance of the feature using Information gain.

Fig. 1 shows the relative importance of the features based on their information gain. The orders of the features are shown as follows in decreasing order, this is a Claim Incurred, Claim Number, Change In out Standing, Estimated Loss, Policy Number, Name of Insured, Net of Recoveries(A-B), Claim paid Gross(A), Net of Recoveries (B) and their corresponding information gain values are 0.176, 0.175, 0.148, 0.115, 0.113, 0.093, 0.075, 0.065, 0.037 respectively. Claim Incurred has highest information gain value. On the contrary, Net of Recoveries (B) has lowest information gain values.

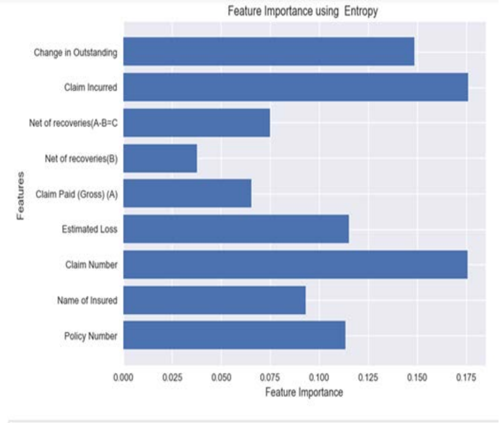

Fig. 1. Relative Feature Importance using Information Gain.

E. Cross Validation Techniques

Machine learning approaches are evaluated using cross validation techniques, it also called rotation estimation. Because the result of cross validation believed that more reliable and less variance to other single train, test split techniques [14] [15]. For this study tenfold cross validation technique was used. 90 % of motor insurance claim data set (58,982 motor insurance claim incurred instances of data sets) used to train the model and 10% of the motor insurance claim data set (6,554 motor insurance claim incurred instances of data sets) used to test the model through iteration.

F. Machine Learning Algorithms

Supervised machine learning algorithms were used to build motor insurance claim status prediction model. For this study, Random Forest (RF) and Support vector machine (SVM) machine learning classifiers were used to build machine learning model. RF classifier consists of many numbers of decision trees as base learners, and each tree train by using random samples of the motor insurance dataset with a replacement which is called bootstrapping. Train all trees by using different samples and take the majority vote for insurance claim status prediction. This process, called Bagging.

Multi class SVM classifier with kernel trick Radial basis function (RBF) and parameter C (cost of penalize misclassification error) with value 1 was used to build motor insurance claim status prediction model. One against all (1AA) approach was used for multi class claim status classification and prediction. In the data set there are five target classes. Therefore, multiple binary class classification was applied using One vs. Rest (OVR) or 1AA approach, because it is efficient to compute and easy to interpret. Five SVM binary classes were built, means that one class vs. the rest classes.

G. Model Performance Evaluation Methods

Machine learning model performance evaluated using different parametric measures, because individual learner gives biased result solutions. Due to this reasons it is useful to measure or evaluate the performance of the algorithm how it is learned from the experience [15]. To evaluate the performance of the model, evaluation metrics were used. For this study, confusion matrices, accuracy, precision, recall and, F-score were used.

Confusion matrix representing as a two dimensional table having predicted values as rows or instances and actual classification values as column. It is not performance measure by its own rather than using other performance metrics with it. These are TP (True positive), TN (True negative), FP (False positive) and FN (False negative) [16]. Accuracy shows the classification problems correct prediction value and calculated as the total number of the model correct prediction divide by all number of data set used for classification. Precision measure the predicted value true and it show how many times the model predicts true.

In the case of Recall the built model identifies the whole relevant examples or instances. F-Measure calculated as by combining the above two methods which is precision and recall as harmonic mean. It is also called F-score, F1- measure. The equation of the above metrics shows as follows.

$$\mathrm{Accuracy(ACC)} = \frac{\mathrm{TN + TP}}{\mathrm{For\ All\ \ Total\ Instances}}$$

$$\mathrm{Pricision(p)} = \frac{\mathrm{TP}}{\mathrm{TP + FP}}$$

$$\mathrm{Recall(R)} = \frac{\mathrm{TP}}{\mathrm{TP + FN}}$$

$$\mathrm{F - score} \ = 2 * \frac{\mathrm{(Recall * Precision)}}{\mathrm{(Recall + Precision)}}$$

IV. Proposed Motor Insurance Claim Status Prediction Model

Fig. 2 shows the proposed model architecture for motor insurance claim status prediction. This architecture has the following components. These are Explanatory data analysis (EDA), Data preprocessing (data cleaning and integration, dimensional reduction, data normalization and encoding), Training and Testing, Evaluate and Model performance comparisons. Fig. 2, shows the detail architecture of the proposed model design.

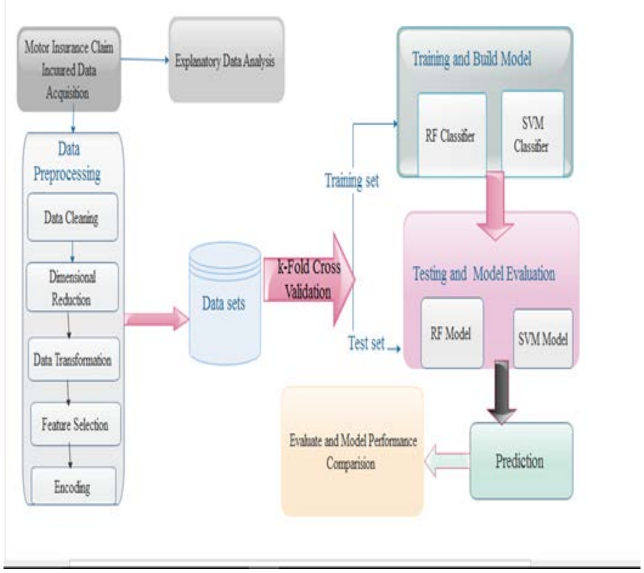

Fig. 2. Architecture of the Proposed Motor Insurance Claim Status Prediction Model.

V. Results and Discussion

A. Evaluation of Result

In machine learning, classification is the most common type of problems [15], because of this there are evaluation metrics, which we used to evaluate the performance of the built machine learning models. For this study, four performance evaluation metrics were used to evaluate the classification performance of the RF, and SVM models using ten – fold cross validation techniques as stated in Section 3F. The data set is split in two parts as training, and testing as it discussed in Section 3D. The two models namely RF and SVM were used, as classifiers. Each classifier is trained and tested. The models obtained, from the training phase were tested by using new motor insurance claim data in addition to, training sets. Accuracy of ten –fold cross validation results were computed by taking the average result of each training set and test sets as demonstrated or illustrated in Table II.

Table II shows the Prediction accuracy of RF and SVM. The RF prediction accuracy in each fold was as follows, 97.45%, 98.94%, 96.99%, 97.03%, 98.39%, 97.07%, 96.73%, 89.42%, 93.17%, and 96.59% on the corresponding experiment 1, experiment 2, experiment 3, experiment 4, experiment 5, experiment 6, experiment 7, experiment 8, experiment 9, and experiment 10 respectively. The lowest percentage result was recorded on experiment 8 (89.42%,) and the highest percentage result was recorded on experiment 2 (98.94%). The average prediction accuracy of RF from those ten experiments is 96.43%.The prediction accuracy of SVM on each fold was 98.96%, 99.19%, 99.11%, 99.40%, 99.63%, 97.22%, 98.10%, 79.18%, 96.45%, and 98.80% on the corresponding experiment 1, experiment 2, experiment 3, experiment 4, experiment 5, experiment 6, experiment 7, experiment 8, experiment 9, and experiment 10 respectively. The lowest percentage score was recorded on experiment 8 (79.18%), similar to RF. The highest percentage score was recorded on experiment 5 (99.63%). The average prediction accuracy of SVM from those ten experiments was 96.60%. Except experiment 8, the accuracy result of the SVM on each experiment was slightly greater than the accuracy result of RF. The performance of the RF, and SVM models clearly illustrated using a bar graph in Fig. 3.

The bar chart in Fig. 3 shows the graphical or visual representation of the above Table I results. The green color represents RF's classification accuracy and the blue color represents the classification accuracy of the SVM's. This bar chart shows the comparison of RF and SVM, how it performs on each fold through iteration.

B. *Classification Result of Models*

The classification performance of the two classifiers (RF and SVM) validated or measured using the test data sets. The results of these classifiers for the test data sets were shown in the Table III and IV, respectively. The column show the actual value and the row show predicted value. The diagonal value of the confusion matrix indicates the correctly classified instances among the test data sets as illustrated below.

Where class, Close, Pending, Notification, Re-open, Settled represent 0, 1,2,3,4, respectively.

The result of each class, TP, FP, FN, TN, accuracy, precision, and F- measure based on RF and SVM models from the confusion matrix report is presented in the Table IV and Table V respectively as shown below.

Table V shows the summary result of RF model. 98.36 % was correctly classified and 1.64 % was misclassified by RF. On the other way, The Precision, Recall and F- measure result of the RF model was 95.15%, 94.71%, and 94.90% respectively. The highest prediction accuracy found for class, re-open, that has 99.83%, and the lowest prediction accuracy for class settled, was 97.34%.

Similarly, Table VI shows the summary of SVM model result of, Accuracy, Precision, Recall and F-measure is 98.17%, 97.22%, 93.80%, and 95.36% respectively and 1.83% was misclassified. The highest prediction accuracy found for class re-open (99.89%) and lowest prediction accuracy was found for class closed (95.94%).

From the above two experimental results, both of the two models have nearly similar prediction accuracy performance. But, RF Model slightly greater than Support vector machine model in terms of accuracy. Both RF and SVM model had the best prediction accuracy of re-open claim status among all other classes of motor insurance claims.

Generally, Random Forest model is slightly better than support vector machine model in both accuracy, and Recall. On the other hand, SVM model better than RF model in both precision and F-measure as summarized in Fig. 4, which shows the comparison of RF and SVM models using the four performance metrics evaluation (Accuracy, Precision, Recall and F- measure).

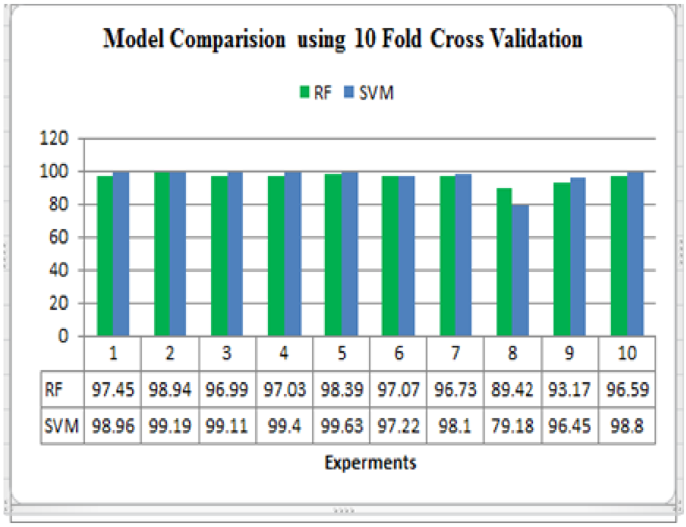

Fig. 3. RF and SVM classification Accuracy Result in Bar Chart.

TABLE II. Test Result for RF and SVM using Each Fold

Experiment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**
Total No. of data sets	65,535										
Accuracy of RF in %	97.45	98.94	96.99	97.03	98.39	97.07	96.73	89.42	93.17	96.59	96.43
Accuracy of SVM in %	98.96	99.19	99.11	99.40	99.63	97.22	98.10	79.18	96.45	98.80	96.60

TABLE III. Confusion Matrix Result for RF Model

	Predicted					
Actual	**Close**	**Notification**	**Pending**	**Reopen**	**Settled**	**Total**
Close	2452	5	34	2	25	2518
Notification	4	685	1	1	1	692
Pending	33	2	798	0	43	876
Re-open	7	0	0	76	1	84
Settled	24	30	50	0	2280	2384
Total	2520	722	883	79	2350	6554

TABLE IV. CONFUSION MATRIX RESULT FOR SVM MODEL

	Predicted					
Actual	Close	Notification	Pending	Re-open	Settled	Total
Close	2393	1	2	1	6	2403
Notification	64	693	4	0	11	772
Pending	84	2	889	0	7	982
Re-open	4	0	0	94	2	100
Settled	104	4	3	0	2186	2297
Total	2649	700	898	95	2212	6554

TABLE V. TP, FP, FN, TN ACCURACY, PRECISION, RECALL, AND F-MEASURE (SCORE) FOR RF MODEL

Class		***TP***	***FP***	***FN***	***TN***	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F-Score (%)***
Closed	0	2452	68	66	3968	97.7955447	97.301587	97.378872	97.3349771
Notification	1	685	37	7	5825	99.328654	94.875346	98.9888439	96.8880576
Pending	2	798	85	78	5593	97.512969	90.373726	91.09589	90.733371
Re-open	3	76	3	8	6467	99.832164	96.202532	90.47619	93.2515336
Settled	4	2280	70	104	4100	97.345133	97.021277	95.637584	96.3247046
Average (%)						**98.3628928**	**95.1548936**	**94.715476**	**94.9065288**

TABLE VI. TP, FP, FN, TN ACCURACY, PRECISION, RECALL, AND F-MEASURE (F- SCORE) FOR SVM MODEL

Class		***TP***	***FP***	***FN***	***TN***	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F-score (%)***
Closed	0	2393	256	10	3895	95.94141	90.335976	99.583854	94.7349474
Notification	1	693	7	79	5775	98.687824	99	89.766839	94.1576084
Pending	2	889	9	93	5563	98.443699	98.997773	90.529532	94.5744684
Re-open	3	94	1	6	6453	99.893195	98.947368	94	96.4102562
Settled	4	2186	26	111	4231	97.909673	98.824593	95.16761	96.9616222
Average (%)						**98.17516**	**97.221142**	**93.809567**	**95.3677806**

According to the above Fig. 4, the result of high value precision in RF and SVM models indicates that, the built model can correctly classify motor insurance claim status and predict the sample data to their corresponding real class.

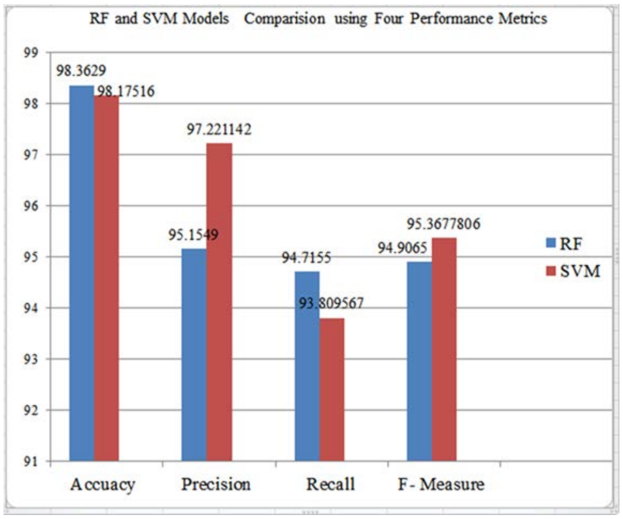

Fig. 4. Models Comparison by using Various Performance Evaluation Metrics.

High recall indicates that many of the data were predicted and high relevant data were selected. Other high value of F-measure shows that best result values are obtained at the precision and recall performance measures. On the contrary, low values of F- measure indicate less value of precision and recall. Generally, the two models give ideal precision- recall results, means that it scores high precision and high recall results.

VI. CONCLUSION

In this study, the potential applicability of machine learning has been implemented and evaluated in the insurance company, specifically for motor insurance claim prediction. This experimental study, which has employed the most powerful, used methodological techniques in machine learning research. So to address the problem, Random forest model and Support vector machine, were used as a predictive model.

In this study, an attempt has been done to design, and implements the model that has a capability of predicting motor insurance claim status. The procedures included data Understanding and explanatory data analysis, data preprocessing), model training, model testing, classification and prediction, and finally comparison of the two built models have done.

The two models built on using 65, 535 instances of motor insurance claim data as input. This input data first needs data understanding and data preparation before to build the two models. The final preprocessed data sets were used for model training and testing. This preprocessed data sets split into two, training set and testing set using K –Fold cross validation with k= 10. Hence, dataset divided in to 10 folds or experiments through iteration. Each fold used as training and testing iteratively, at least each fold used once as testing set. Finally the average score for each fold was taken. The performances of the two classifiers were evaluated by using four metrics (Accuracy, Precision, Recall and F-measure). Therefore, the experimental result shows that the two classifiers score an overall accuracy of 98.36929% and 98.17516%, correctly classified by the two models respectively.

Generally, the performance of the model was evaluated with four metrics (Accuracy, Precision, Recall, and F-measure). The developed motor insurance claim status prediction models have best prediction accuracy, and the two models have promising prediction accuracy. RF model prediction accuracy is slightly better than SVM model in the insurance domain specifically in motor insurance.

VII. FUTURE WORK

In this study, a good result was achieved in predicting motor insurance claim status. But, it was not possible to implement all machine learning classification algorithms, because of this the researchers propose extending this study with other machine learning algorithms, and build hybrid machine learning model using graphical user interface design to apply in the real world insurance companies.

REFERENCES

[1] Hailu Zeleke ., “Insurance in Ethiopia: Historical Development, Present Status and Future Challenges .,” vol. 1, no. 1, p. 308, 2009.

[2] K. P. M. L. P. W. and M. C. W. Depa, “A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims,” Eur. Int. J. Sci. Technol., vol. 5, no. 1, pp. 47–54, 2016.

[3] A. C. Yeo, K. A. Smith, R. J. Willis, and M. Brooks, “Clustering Technique for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry,” Int. J. Intell. Syst. Accounting, Financ. Manag., no. November 1999, pp. 39–50, 2001.

[4] M. C. Wijegunasekara and Weerasingheand M.C. Wijegunasekara , “A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims,” vol. 5, no. 1, pp. 47–54, 2016.

[5] K. A. Smith, R. J. Willis, M. Brooks, K. A. Smith, R. J. Willis, and M. Brooks, “An analysis of customer retention and insurance claim patterns using data mining : a case study,” J. Oper. Res. Soc. ISSN, no. 5682, pp. 1476–9360, 2017.

[6] W. Lin, Z. Wu, L. Lin, A. Wen, and J. I. N. Li, “An Ensemble Random Forest Algorithm for Insurance Big Data Analysis,” IEEE Acess, vol. 5, 2017.

[7] P. Bharal and A. Halfon, “Making Sense of Big Data in Insurance,” ACORD and MarkLogic, 2013.

[8] L. Wang and C. A. Alexander, “Big Data : Infrastructure , technology progress and challenges,” J. Data Manaagement Comput. Sci. Vol., vol. 2, no. 1, pp. 1–6, 2015.

[9] A. L. Heureux and M. Grolinger, Katarina and Caprtz, “Machine Learning With Big Data : Challenges and Approaches,” IEEE Access, vol. 5, pp. 7776–7797, 2017.

[10] A. S. Alshamsi and A. Ain, “Predicting Car Insurance Policies Using Random Forest,” IEE, pp. 128–132, 2014.

[11] Endalew Alamir, Teklu Urgessa, T. GopiKrishna and Ellappan V, “Application of Machine Learning with Big Data Analytics in the Insurance,” vol. 11, no. 12, pp. 1064–1073, 2020.

[12] A. S. Alshamsi and A. Ain, “Predicting Car Insurance Policies Using Random Forest,” pp. 128–132, 2014.

[13] T. Kavipriya and N. Kumar, “A Study on Machine Learning Algorithms for Big Data Analytics,” IOSR J. Eng., no. Iccids, pp. 40–46, 2018.

[14] A. C. Tan and D. Gilbert, “An empirical comparison of supervised machine learning techniques in bioinformatics,” First Asia Pacific Bioinforma. Conf. (APBC 2003), vol. 19, no. Apbc, 2003.

[15] J. Brownlee, Machine Learning Mastery with python, V1.4. 2016.

[16] R. J. Kate and A. M. Swartz, “Assessment of various supervised learning algorithms using different performance metrics Assessment of various supervised learning algorithms using different performance metrics,” IOP Conf. Ser. Mater. Sci. Eng., 2017.

[17] C. Using, S. Vector, F. K. C-means, Z. Rustam, and F. Yaurita, “Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance .,” *J. Phys. Conf. Ser. Pap.*, 2018.

[18] N. K. Frempong, N. Nicholas, and M. A. Boateng, “Decision Tree as a Predictive Modeling Tool for Auto Insurance Claims,” *Int. J. Stat. Appl.*, vol. 7, pp. 111–120, 2017.