ICIC Express Letters
Part B: Applications
Volume **16**, Number **8**, August **2025**

ICIC International ©2025 ISSN 2185-2766
pp. **805–812**

FRAUDULENT HEALTH INSURANCE CLAIMS DETECTION USING MACHINE LEARNING: A CASE STUDY IN THAILAND

Kanyanut Homsapaya[1] and Nittaya Thonghnunui[2,*]

[1]Department of Computer Science and Information
Kasertsart University, Sriracha Campus
199 Moo 6 Sukhumvit Road, Thung Sukhla, Si Racha, Chon Buri 20230, Thailand
kanyanut.ho@ku.th

[2]Department of Statistics
Rambhai Barni Rajabhat University
41 M.5 Tambol.Tachang, Aumpher.Muang, Chanthaburi 22000, Thailand
*Corresponding author: nittaya.th@rbru.ac.th

Received October 2024; accepted January 2025

Abstract. *Fraudulent healthcare insurance claims cause an ongoing critical problem that affects the growth rate of healthcare services in many countries because of increased medical costs since companies usually pay the claims amount when they should not have to. Based on Health Insurance's medical claims submission data from 2019 to 2021, fraudulent claims have significantly increased over time. Moreover, the large volume of claims submissions on a daily basis (approximately 2.5 million per day) makes human reviews difficult. Currently, healthcare organizations have traditionally depended on manual review processes by claims assessors. However, these processes are time-consuming and often do not identify fraudulent activity due to the high volume of claims submissions each day and the service level agreement target. Thus, an automated claims assessment system to reduce the inefficient review process by humans is essential. Herein, we present an automated fraudulent claims detection model based on machine learning. This has a large impact on the claims assessment process by successfully integrating digital transformation to reduce costs, improve accuracy, and be less laborious. Our model based on supervised machine learning was implemented and evaluated for a health insurance service provider on approximately 800,000 insurance claims records from hospitals within the past three years. In the preprocessing step, we manage multiple tasks such as discretization to classify the data into discrete values and applying the Min-Max scalar for feature scaling and balancing. For the classifier model, Decision Tree, Random Forest, Naïve Bayes, and XGBoost are used to evaluate the performance of model Accuracy, Precision, Recall, and area under the receiver operating characteristic curve. We compared the model performance with and without preprocessed data to prove that the selected preprocessing methods could significantly improve the model performance. We designed a data pipeline comprising data preprocessing on the premises and ran machine-learning models on Python (AWS Cloud), after which we compared the classification rates using various algorithms; the best results were obtained using Random Forest.*

Keywords: Healthcare, Anomaly detection, Fraud claims detection, Machine learning, Health claim insurance

1. **Introduction and Literature Review.** Although good health is the most valued prerequisite for human life, health insurance protects subscribers from a potentially large financial burden due to unexpected healthcare problems. Moreover, having health insurance allows affordable access to quality healthcare providers and doctors, and is sought by people in Thailand who can afford and are willing to pay the policy premium amount. According to the Thai General Insurance Association, fraudulent claims are becoming an increasingly serious problem.

DOI: 10.24507/icicelb.16.08.805

Healthcare fraud is a crime in the health insurance domain [1] and 9% of healthcare services are wasted because of it [2]. Consequentially, fraudulent claims damage the revenue and profit of insurance companies [3]. Before the era of digital transformation, the claims operation process handled many thousands of claims via paper submissions from providers or claimants, which makes it time-consuming to pay attention to each claim [4]. Moreover, the high cost of fraudulent claims investigations is another concern. As a result, insurance companies have attempted to deal with claims costs and time by exploiting artificial intelligence and machine learning to detect fraudulent claims. By analyzing historical data (i.e., past behavior) of the claims, machine-learning algorithms can learn to discriminate between fraudulent and non-fraudulent activity and use this information for further investigations.

Kirlidog and Asuk [5] stated that the business fraud rule based on claims auditors can be easily applied to capturing fraudulent claims data from a database using SQL. Afterward, insurance experts can review them because only suspicious transactions have been extracted. The fraudulent data concerns not only individual claimants but also hospital providers and pharmacies. Capelleveen et al. [6] stated that private health insurance in Chile lose the huge amount of revenue from fraud and abuse medical claims increasingly by year. Manually approval claim process is extremely expensive procedure due to lacking efficient results which have responsibility to own the tasks. Then, the proposed solution is employed multilayer perceptron neural network (MLP) and the outcome of fraud detection application that can be successfully replaced the manual process to the concrete business problem to combat against with fraud claim data. Lakhan et al. [7] demonstrated the costly problem of claim approval human process, and the effect is getting higher with loss ratio. The paper stated two points of problems that are data scarcity of insurance claims from private companies and imbalanced datasets. They proposed a new technique which related a distributed and concealment-preserving federated solution by employing the adjusted random forest model and the outcome improved prediction accuracy rate and efficiency. The discovery of this paper [8,9] shows that fraud claim detection is applied with unsupervised outlier techniques to automatically find the suspicious fraud pattern that received from claim submission process. The encouragement to initialize the automatic machine-learning model is that insurance company suffered from a large amount of claim process without time efficiency and good analysis in suspicious fraud claim data. The results can help to flag anomaly detection or suspicious claim for aiding claim expert to investigate further step and finally, reveal the fraud case.

This study focuses to enlighten on the significance of using ML in automated detecting fraudulent cases on claim submission application. Consequently, there is a strong need to change traditional manual processes and employ machine learning to identify fraudulent cases within the extensive volume of patient claim submissions.

The methods explored in the reviewed literature using machine learning involved the comparison of Decision Tree, Random Forest, Naïve Bayes, XGBoost, and KNN [10,11]. Essential metrics for evaluating the performance of these models included the confusion matrix, Accuracy, Precision, Recall, F1, and Receiver Operating Characteristic (ROC). Additionally, we concentrated on the most accurate and simplest approach to reduce fraudulent claims and focused on changing from the manual approval process to automated detection by using supervised machine-learning techniques. First, claims submission data from 2019 to 2021 provided by the Health Insurance Company in Thailand were digitally transformed and then used in the analysis. Three key performance indicators (KPIs) were considered in this approach: time efficiency, cost savings, and accuracy.

The remainder of this paper is structured as follows. Section 2 introduces claim data set and research stage. Section 3 presents our methodological approach for detecting insurance fraud using supervised machine learning and evaluation approach. Section 4 presents the results from our evaluation approach. Section 5 concludes the paper.

2. **Claim Data Set and Research Stage.** In this phase, data were transferred from the on-premises health claims database to the AWS S3 service. An initial exploratory analysis of the data was conducted by using the Pandas library. The total number of claims in 2019, 2020, and 2021 comprised 205,000, 220,000, and 245,000 outpatient department (OPD) and in-patient department (IPD) records, respectively.

2.1. **Business understanding and gap analysis.** The meetings were held with expert domains several times comprising medical experts, nurses, claim operation team who explained the main approval claim process. This involved gathering details of the claim operation process, understanding the insurance technical terms such as diagnosis code groups ICD9 and ICD10, OPD, and IPD, uncovering the business problems that needed to be solved, and determining how many KPIs the digital transformation team needed to complete the task successfully.

2.2. **Data collection and preparation.** The fraudulent claims detection model was considered to be a classification problem in which supervised machine learning is used to identify input features and predict which class the features fall into. Subsequently, the data labels were mandated and separated into two classes: fraudulent and non-fraudulent by using pattern identification based on the experiences of fraudulent claims supervisors. For our claim's datasets, 33 attributes were considered for assignment and the data were then classified into fraudulent and non-fraudulent. The datasets contained 898,045 claims transaction records. The numbers of fraudulent and non-fraudulent cases were 878,680 and 19,365, respectively.

2.3. **Data exploration.** As can be seen in Table 1, the approved claims amount has increased each year by around 9% and 34% for 2020 and 2021 compared to 2019, respectively. The number of claims has increased each year by around 7.14% and 15.55% for 2020 and 2021 compared to 2019, sequentially.

Table 1. Comparison approved claim data 2019-2021

Type	Year		
	2019	2020	2021
Approved claim amount (billions)	0.95	1.04	1.40
No. of claims (thousands)	210	225	260

Table 2. Comparison of fraud no. of claims 2019-2021 (thousands)

Month	Fraud		
	2019	2020	2021
1	16	30	20
2	14.5	23	35
3	15	21.5	22
4	14	15	19
5	17	17.5	21
6	15	18	22
7	17.5	20	19
8	18	22	17
9	18	21	23
10	21	17	19.5
11	20	10	22.5
12	23	18	21

As can be seen in Table 2, the number of claims has increased each year by around 7.14% and 15.55% for 2020 and 2021 compared to 2019, respectively.

According to the results, Table 3 illustrates the fraudulent and non-fraudulent claims amounts by month and shows comparison claim amounts of fraud and non-fraud classes in 2019-2021. In 2019, the range of fraud/non-fraud ratios was found between 6% to 13% for the whole year. Secondly, for the monthly claims amount in 2020, the fraudulent claims varied from 9% to 34% over the year. In 2021, the results show that the monthly claims amounts in which the fraudulent claims varied from 5% to 10% over the year.

TABLE 3. Comparison of fraud and non-fraud approved claim amount 2019-2021 (millions)

Month	Fraud			Non-fraud		
	2019	2020	2021	2019	2020	2021
1	5.75	15.58	6.59	58.44	108.36	84.27
2	5.05	8.93	9.27	61.24	84.81	133.44
3	4.33	10.48	5.64	72.34	80.81	104.65
4	7.09	10.12	7.01	58.80	70.58	83.48
5	6.61	22.8	6.87	74.45	67.80	113.03
6	8.26	10.68	6.98	63.66	81.24	106.40
7	6.87	9.62	6.28	65.13	77.96	120.59
8	9.28	8.93	11.9	69.19	84.77	120.40
9	9.33	6.65	7.35	75.75	85.00	126.44
10	10.15	4.25	7.16	91.12	65.03	104.70
11	6.53	4.96	6.47	80.99	56.17	116.00
12	9.72	6.49	6.41	85.67	58.32	104.70

3. **Methodology.**

3.1. **Data cleansing and transformation.** This phase improved the quality of the data for the subsequent analysis. Problems arose when merging many data sources due to a lack of data consistency. According to Thornton et al. [12], data cleansing is of foremost importance before applying algorithms to datasets to prevent misinterpretation and wrong outcomes. These were our preprocessing steps in the present study. First, rows were deleted if they contained missing values, noise, or inconsistencies. Second, the data were categorically encoded as integers for Gender, Benefit_Head, and Benefit_Type. In another investigation conducted by Ambarwari et al. [13], it was demonstrated that data scaling methods like Min-Max normalization and standardization exert notable influences on data analysis.

3.2. **The features used in the machine-learning model.** Feature selection is the process by which relevant features that reduce the model complexity are selected. Filtering techniques include forward selection starting with zero features and adding one with each iteration and backward elimination starting with all of the available features and removing one with each iteration and evaluating the performance of the classifier afterward. However, since both of these are limited, sequential floating search algorithms that provide better classification performance have been proposed in 1994 by Pudil et al. [14]. Moreover, to enhance the quality of selected features in each iteration, a genetic algorithm is employed and integrated with sequential floating search. A mutual information function is then applied to selecting relevant and high-quality features, contributing to an enhancement in classification accuracy [15].

3.3. **Class balancing.** The fraudulent and non-fraudulent datasets were found to be unbalanced data because of differences in classification, which could have led to bias in the model. The imbalance ratio was 42 : 1. Simply explained, the imbalance problem is caused by the claims dataset predominantly comprising non-fraudulent cases, and the results of the evaluation metrics are predisposed toward the center of the majority of cases. Consequently, a classifier's success at predicting the most frequent cases may fail to accurately predict the correct class of minority cases, which devalues the main goal performance of the model [16]. Regarding this research problem, SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods, so it was applied to solving this issue, the objective being to balance the two datasets by class. SMOTE oversamples the minority class to generate virtual training records via linear interpolation using the k nearest neighbor method to select the random nearest neighbor, thereby randomly creating an artificial instance in feature space [17].

The data were divided into training and testing datasets with a ratio of 70 : 30 and 10 cross-validation processes. To evaluate the reliability of the classifiers, the following assessments were used: confusion matrices, which demonstrate the accuracy by statistically comprehending actual positive and negative values through false positive and negative classifications. Although the ubiquitous accuracy metric can be used to measure the performance of a model, it is not applicable in our case due to the minority class problem. Subsequently, the Area under the Receiver Operator (AUC), Precision, Recall, and F1 were used as evaluation metrics. AUC can be used to summarize the classifier performance into only one measure and normally helps to make the decision about which classifier offers the most quality. Thus, an AUC value close to 1 means a good classification performance whereas less than 0.5 demonstrates that the classifier performance is poor. Thus, the higher the AUC, the greater the classifying model. In addition, the ROC can be utilized to assess the overall classification performance by plotting the relationship between the true positive rate and the false positive rate to provide the decision value. The F1 score combines precision and recall using their harmonic mean. Increasing the threshold for the F1 score implies simultaneously maximizing for both precision and recall. Thus, the F1 score will only be high if both precision and recall are high, indicating a good balance as a safeguard of both measures [18].

Based on confusion matrix, the mathematical formulae for the evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4}$$

4. **Results and Discussion.** As can be seen in Figure 1, the presented graph demonstrates ROC curves for individual classification methods and corresponding SMOTE techniques utilized in this study. It effectively demonstrates the classifier performance across all machine learning algorithms explored in this paper. ROC curve provides valuable insights into a classification system's ability to differentiate between two classes. The most superior curve is associated with Random Forest, represented by the line with square, while the least satisfaction curve is associated with Naïve Bayes. Subsequently, the second position is occupied by the Decision Tree, highlighted in the line with circle, the third and the fourth are displayed by XGBoost and KNN.

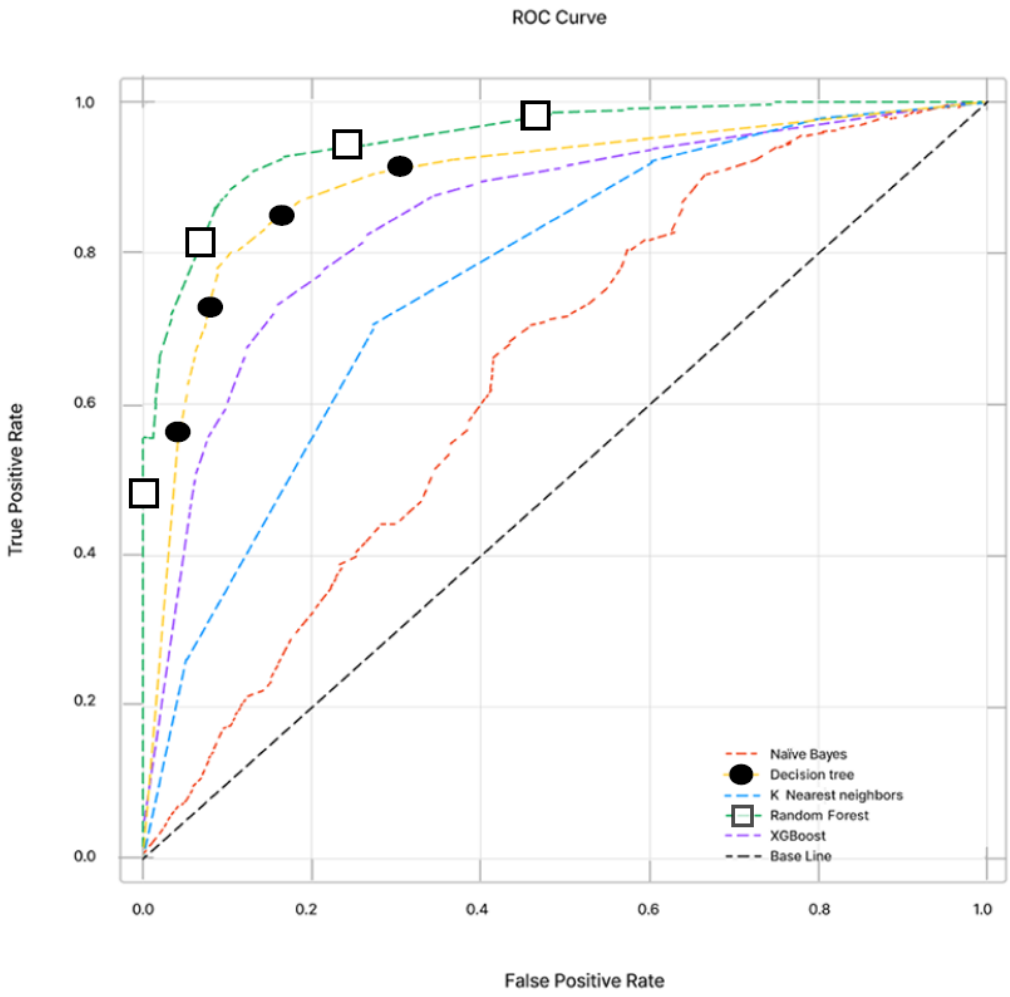

FIGURE 1. Comparing ROC curves for all the models

TABLE 4. Metric results of 10 cross-validation

Algorithm	Accuracy rate	Precision	Recall	F1
Decision Tree	0.83	0.89	0.91	0.90
Random Forest	0.85	0.90	0.92	0.92
Naïve Bayes	0.67	0.75	0.77	0.76
KNN	0.75	0.81	0.84	0.82
XGBoost	0.80	0.86	0.89	0.88

To obtain the performance results from Table 4 by using the Scikit-learn, Pandas and Numpy machine-learning library, we computed the classification rate by comparing algorithms with 10 cross-validation cycles. We considered real insurance claims datasets in this study and executed a sample of 30,000 customers with corresponding attributes from the database and the evaluation was conducted on 230,000 unseen health insurance claims preprocessed via the previous task. The results provided by using Random Forest provided the best performance, which is 85%, followed by Decision Tree (83%, which is very close to the Random Forest result). Next, the Naïve Bayes model demonstrates a comparatively lower accuracy performance than the KNN model (67% versus 75%). The findings indicate that model performance can be enhanced through feature engineering and feature selection utilizing both floating and genetic algorithms. Additionally, the performances when using 5 and 10 cross-validation cycles were almost the same.

Insurance companies want to detect fraudulent claims (FN), so Recall (TPR) is a very significant measure since we do not want to miss any fraud transactions. On the other hand, FP also offers crucial information to retrieve because the caregiver must contact the insurer to determine whether the case is real or not. Moreover, high Precision and Recall values are needed for the quality classifier. Consequently, Random Forest is a suitable model for our claim data sets. This study should be created from the foundation for future investigations into detecting fraud. It concentrates on the primary factors influencing healthcare fraud in this specific context, offering valuable insights into fraudulent practices within the Thailand healthcare system. The utilization of machine learning is proposed to automate the fraud detection process, minimizing human intervention, and mitigating potential subjectivity associated with manual procedures. This not only enhances the precision of fraud detection but also saves processing time. It is crucial to note that the predictive accuracy of the model is contingent on the quality of the acquired data; the model's effectiveness is directly tied to the quality of the data it employs. Furthermore, continuous retraining and updates are imperative to address any concept drift that may arise over time.

5. **Conclusions.** A method for applying a supervised learning classifier to using actual healthcare insurance claims data comprising approximately 800,000 transactions is presented herein. A domain expert team labeled the claims as fraudulent and 16 genuine and provided information about how to detect suspicious behavior. EDA provided excellent insightful information while a data analyst interpreted the data in the analytic graph. Besides, data cleansing, feature selection, feature engineering, and applying SMOTE were conducted to solve the imbalance problem before model creation. In the construction of the model, classification algorithms Decision Tree, Random Forest, XGBoost, KNN and Naïve Bayes were employed to discover the best classifier algorithm for the fraudulent claims detection system based on real insurance claims datasets. In the future, the fraudulent claims detection method could be combined with others to improve accuracy. Finally, we communicated the final model to the stakeholders and deployed it to process claims data in the production system. Compared to the prior claims assessment process, our successful KPIs were more efficient claims detection and reduced processing time for claims assessments. Thus, applying AI to this real-life situation greatly improved their daily work routine for claims assessment. It can be concluded that the model improves healthcare insurance quality and yields positive economic results. Future research will be focused on improving machine learning methods to detect health insurance claims fraud using Precision-Recall curves as a main metric to evaluate the performance of the classifiers.

Acknowledgement. The authors are grateful for financial support from the Cigna Insurance Company. We particularly thank you for Project mentors, Dr. Benny Wong and Miss Napha, for valuable insight and important knowledge feedback.

REFERENCES

[1] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri and M. Arab, Using data mining to detect health care fraud and abuse: A review of literature, *Global Journal of Health Science*, vol.7, no.1, 194, 2015.

[2] Z. X. Chen, L. Hohmann, B. Banjara, Y. Zhao, K. Diggs and S. C. Westrick, Recommendations to protect patients and health care practices from medicare and medicaid fraud, *Journal of the American Pharmacists Association*, vol.60, no.6, pp.e60-e65, 2020.

[3] P. N. Jyothi, M. V. Ramana and S. Suresh, A comparative analysis for identifying the fraudulent healthcare claims, *Grenze International Journal of Engineering and Technology*, 2023.

[4] F. Aslam, A. I. Hunjra, Z. Ftiti, W. Louhichi and T. Shams, Insurance fraud detection: Evidence from artificial intelligence and machine learning, *Research in International Business and Finance*, vol.62, 101744, 2022.

[5] M. Kirlidog and C. Asuk, A fraud detection approach with data mining in health insurance, *Procedia-Social and Behavioral Sciences*, vol.62, pp.989-994, 2012.
[6] G. van Capelleveen, M. Poel, R. Mueller, D. Thornton and J. Hillegersberg, Outlier detection in healthcare fraud: A case study in the medicaid dental domain, *International Journal of Accounting Information Systems*, vol.21, pp.18-31, 2016.
[7] A. Lakhan, M. A. Mohammed, J. Nedoma, R. Martinek, P. Tiwari, A. Vidyarthi, A. Alkhayyat, W. Wang and P. Tiwari, Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare, *IEEE Journal of Biomedical and Health Informatics*, vol.27, no.2, pp.664-672, 2022.
[8] K. Nian, H. Zhang, A. Tayal, T. Coleman and Y. Li, Auto insurance fraud detection using unsupervised spectral ranking for anomaly, *The Journal of Finance and Data Science*, vol.2, no.1, pp.58-75, 2016.
[9] A. Bhaskar et al., An intelligent unsupervised technique for fraud detection in health care systems, *Intelligent Decision Technologies*, vol.15, no.1, pp.127-139, 2021.
[10] J. O. Awoyemi, A. O. Adetunmbi and S. Oluwadare, Credit card fraud detection using machine learning techniques: A comparative analysis, *2017 International Conference on Computing Networking and Informatics (ICCNI)*, 2017.
[11] E. Nabrawi and A. Alanazi, Fraud detection in healthcare insurance claims using machine learning, *Risks*, vol.11, no.9, 160, 2023.
[12] D. Thornton, G. van Capelleveen, M. Poel, J. Hillegersberg and R. Mueller, Outlier-based health insurance fraud detection for US medicaid data, *Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS 2014)*, vol.2, pp.684-694, 2014.
[13] A. Ambarwari, Q. J. Adrian and Y. Herdiyeni, Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol.4, no.1, pp.117-122, 2020.
[14] P. Pudil, J. Novovičová and J. Kittler, Floating search methods in feature selection, *Pattern Recognition Letters*, vol.15, no.11, pp.1119-1125, 1994.
[15] K. Homsapaya and O. Sornil, Modified floating search feature selection based on genetic algorithm, *MATEC Web of Conferences*, vol.164, 2018.
[16] P. Gnip, L. Vokorokos and P. Drotár, Selective oversampling approach for strongly imbalanced data, *PeerJ Computer Science*, vol.7, e604, 2021.
[17] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol.16, pp.321-357, 2002.
[18] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters*, vol.27, no.8, pp.861-874, 2006.