MINISTRY
OF EDUCATION AND SCIENCE
OF UKRAINE

ACADEMY
OF TECHNICAL SCIENCES OF UKRAINE

DEPARTMENT OF
INFORMATION
TECHNOLOGY

CONNECTIVE
technologies

# 2022 International Conference on Innovative Solutions in Software Engineering (ICISSE)

Conference Proceedings

November 29-30, 2022
Ivano-Frankivsk, Ukraine

*Issued by the decision of the Academic Board of the Faculty of Mathematics and Computer Science (protocol #15/2022)*



***Editors***
**Mykola Kuz** and **Mykola Kozlenko**
Department of Information Technology
Faculty of Mathematics and Computer Science
Vasyl Stefanyk Precarpatian National University
57 Shevchenka Street
Administration Building, Office #319
Ivano-Frankivsk, 76018, Ukraine
Tel: +380 (342) 59-60-58
E-mail: icisse@pnu.edu.ua



**2022 International Conference on Innovative Solutions in Software Engineering (ICISSE),** Mykola Kuz and Mykola Kozlenko Eds., Ivano-Frankivsk, Ukraine: Vasyl Stefanyk Precarpathian National University, 2022, 317 p. [E-book]

**ISBN 978-966-640-534-3**
**DOI 10.5281/zenodo.7502536**

The high-quality refereed papers appearing in this book compose the proceedings of the 2022 International Conference on Innovative Solutions in Software Engineering (ICISSE). ICISSE is the forum organized by the Department of Information Technology of the Vasyl Stefanyk Precarpathian National University. ICISSE-2022 is held from November 29 to 30, 2022 in Ivano-Frankivsk, Ukraine. The event is intended to bring together researchers, scientists, and engineers to discuss experimental and theoretical results, mainly in the area of software engineering. The conference also covers topics related to computer science, computer engineering, systems analysis, cybersecurity, information systems and technology, industrial automation, electronics, metrology, micro and nanosystems, telecommunications, radio frequency engineering, IT entrepreneurship, and IT education. The current edition of the event is dedicated to the 50th birth anniversary of famous Ukrainian scientist working in information technology, Founder and first Head of the Department of Information Technology, Vice-Rector for Research of the university, Professor Dr. Pavlo Fedoruk, who passed away in 2013. The conference is included into the list of scientific conferences scheduled for 2022 by the Ministry of Education and Science of Ukraine.

Recommended paper citation template:
J. K. Author, "Decapitalized title of the paper," *2022 International Conference on Innovative Solutions in Software Engineering (ICISSE)*, Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine, Nov. 29-30, 2022, pp. xxx-xxx.

**ISBN 978-966-640-534-3**

# COMMITTEES

## General Chair:
Prof. Dr. Mykola Kuz

## Technical Program Committee Chair:
Assoc. Prof. PhD Mykola Kozlenko

## Local Organizing Committee Chair:
Mariia Dutchak

## Publicication Chair:
Assoc. Prof. PhD Mykola Kozlenko

# Preface

The 2022 International Conference on Innovative Solutions in Software Engineering dedicated to the 50th anniversary of an outstanding researcher and teacher, the founder of the Department of Information Technology at Vasyl Stefanyk Precarpathian University, Prof. Dr. Pavlo Fedoruk was held on November 29-30, 2022 in Ivano-Frankivsk, Ukraine.

Vasyl Stefanyk Precarpathian National University has more than 80 years of history. Its inspirational success story was written by people who devoted themselves to science, education, and culture. Among the prominent figures of the University, the name of the outstanding modern Ukrainian scientist Pavlo Fedoruk shone as a bright star. Now he would have celebrated his 50th birthday.

Pavlo Fedoruk graduated from Precarpathian University in 1994. Immediately after graduating, he was offered the position at the Information Department of the University. He proved himself to be a professional in the field of information technology, a talented person with progressive thinking. He was offered the position as a head of the Information Department in October 1994. Since 2002, Pavlo Fedoruk worked as a director of the Center for Information Technology. He also founded the Center of Distance Learning. While performing his official duties, he was simultaneously involved in research. He was a hard worker and independent thinking researcher who worked effectively on telecommunication technology research.

He studied as a PhD student in the period from 1999 to 2002. He received his PhD from the Institute of Mathematical Machines and Systems of the National Academy of Sciences of Ukraine. The topic of the PhD thesis was related to the systems of distance learning and knowledge assessment based on Internet technology. During his work at the University, Pavlo Fedoruk published more than 120 research papers and two monographs. He developed a course on Informatics for students of the Faculty of Economics and published the textbook "Computer for Economists". Also he developed courses on artificial intelligence and intelligent systems for students of the Faculty of Mathematics and Computer Science. In the period from 2006 to 2008, he was a scholarship holder of the Cabinet of Ministers of Ukraine for young scientists. In 2010, he was successfully awarded with the degree of Doctor of Technical Science. The topic of the thesis was related to adaptive systems of distance learning.

Since 2007, he was a member of the program committee and a session chairman at the international scientific and practical conferences "Education based on web technologies" and "Advanced computer technology in education". In 2007, he participated in the International Referee Committee as a representative from Ukraine. In 2008, he was an expert of the Scientific and Technical Council of the Ministry of Education and Science of Ukraine at the Informatics and Cybernetics section. He was a member of specialized academic councils at the Institute of Mathematical Machines and Systems and Ternopil National Economic University. He supervised four research projects in the field of distance learning. In 2009, he was awarded the Tersenov Regional Prize for Innovation from the Society of Inventors and Innovators of Ukraine. In 2010, he was recognized as a Famous Scientist of the Year for significant achievements in the field of research and education and for personal contribution. In 2011, he was awarded an honorary certificate from the Cabinet of Ministers of Ukraine for his significant personal contribution to science.

In 2012, he was promoted to the position of vice-rector for research of the University. His views were based on a broad knowledge and interests. He was a founder of many scientific, social, and entrepreneurial projects. He was one of the founders of the "Firtka" Local News Agency. He was also the President of the Ivano-Frankivsk Regional Martial Arts Federation.

The authors of this collection devoted their papers to the memory of Prof. Dr. Pavlo Fedoruk.

## TABLE OF CONTENTS

# Selection of data preprocessing technique for imbalanced dataset

Olga Solovei

*Department of Information Technology and Applied mathematics*
*Kyiv National University of Construction and Architecture,*
Ukraine

***Abstract*—In many domains datasets are imbalance, most often the problem is visible in datasets with a medicine diagnoses' prediction; credit's approval; surveys. The results of machine learning algorithms for imbalanced dataset are biased to majority class, whereas the correct predictions of the smaller class somethings could be most important. The goal of the work is to empirically study which data preprocessing technique: feature selection or oversampling to be applied before to train a classification learning algorithm.**

***Keywords—Oversampling, Feature Selection, Area under ROC curve, Random Forest classifier.***

## I. INTRODUCTION

In study [1] to improve classification models' results for imbalance datasets are proposed the following data preprocessing strategies: 1) oversampling - the smaller class is oversampled to match the size of the major class; 2) down-sizing – the number of samples in the major class is reduced; 3) learning by recognition - to use unsupervised machine learning algorithms and ignore target class in dataset. From the mentioned three strategies the preference is given to oversampling technique as it makes balance distribution between classes which improves that classifier's performance. However, in study [2] can be found that the most widely used oversampling method - SMOTE is good for a highly dimensional imbalance dataset only when k-nearest neighbors classifier is used and most predictive features are selected.

Because the results of the mentioned studies look controversial more empirical tests are required to understand the optimal approach to preprocessing technique selection.

## II. RELATED WORK

In work [3] was proposed to use only feature selection technique for a highly dimensional imbalance gene dataset without comparing its efficiency to the results of the oversampling or down-sizing, therefore the concluded results of the study are not obvious.

To eliminate that gap, in the current work, will be conducted experiments to evaluate the efficiency of data preprocessing techniques: 1) a feature selection; 2) an oversampling by SMOTE; 3) a combination of both.

## III. METHODOLOGY

Our datasets are from UCI Machine Learning Repository: "adult", "coli" and "songs" are highly dimensional with imbalance ration between the minority to the majority class: 0.3, 0.06, 0.16 correspondingly. The features' count per each dataset is 81, 86 and 58 and samples' count per each dataset is 48842, 9822, 1409 correspondingly. In the experiments are included the steps: 1) create for each dataset two pairs of the train and test data using different data preprocessing techniques: pairset1 – the dataset is left imbalance but the most predictive features are selected by applying mutual information method [4]; pairset2 - the dataset is balanced by SMOTE method so that a minority class 100% matches the majority and the most predictive features are selected; pairset3 - the dataset is balanced by SMOTE method but all features are included. 2) train Random Forest

(RF) algorithm and evaluate the prediction quality using metrics: area under ROC (AUC), precision, recall and f1-score on the test subset; 4) compare performance's results based on the metrics' values and ROC curves' graphs.

## IV. RESULTS

Random Forest classifier's performance metrics for each dataset are recorded in table 1. ROC convex hull graphs presented on fig 1.

Table 1. RF algorithm's performance results and training time

| dataset | Pairset1 | | | | Pairset2 | | | | Pairset3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC, % | Precision, % | Recall, % | f1 score, % | AUC, % | Precision, % | Recall, % | f1 score, % | AUC, % | Precision, % | Recall, % | f1 score, % |
| adult | 84 | 64 | 51 | 56 | 93 | 84 | 86 | 85 | 93 | 85 | 86 | 85 |
| coli | 65 | 52 | 51 | 51 | 99 | 96 | 95 | 96 | 99 | 97 | 95 | 96 |
| songs | 47 | 57 | 54 | 55 | 85 | 81 | 80 | 80 | 90 | 83 | 83 | 83 |

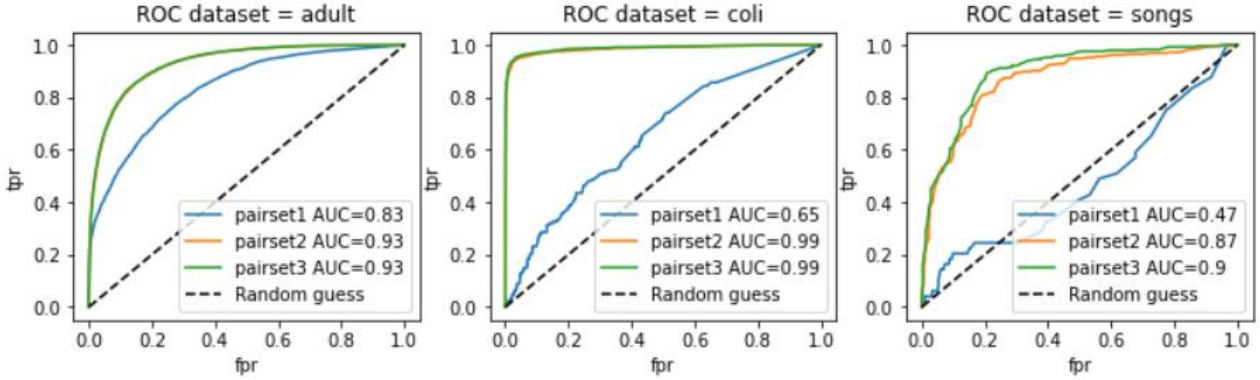


Figure 1. ROC curves of RF classifier trained on pairset1, partset2 and pairset3.

## V. DISCUSSION

The results from table 1 can be interpreted as the following: for highly dimensional datasets an ability of RF classifier to predict the target class is better when a dataset is balanced (pairset2, pairset3) compared to the performance of RF on the imbalance dataset with the most predictive features (pairset1). However, for dataset = "coli" the performance of RF on the balance dataset with all features (pairset3) is better compared to the performance of RF on the balance dataset with the most predictive features are selected (pairset2).

The same results are visible on ROCs convex hull graphs from fig 1: the "more northwest" ROC curve is seen for pairset2 and pairset3 on datasets: "adult" and "coli" and the "more northwest" ROC curve is seen for pairset3 on dataset: "songs".

## VI. FUTURE RESEARCH

It is worth to further investigate: 1) the patterns in the dataset when applying only feature selection technique can guarantee the high model's performance; 2) how model's performance is affected when down-sizing technique together with feature selection are applied on the data preprocessing stage.

## VII. CONCLUSION

Random Forest classifier's performance is better when SMOTE method of the oversampling technique is applied compared to the performance's results received by applying dataset's dimension reduction by selecting the most predictive features using mutual information method.

## REFERENCES


[1] N. Japkowicz. “The class Imbalance Problem: Significance and Strategies,” *AAAI conference of Artificial Intelligence*,2000,pp.10-15.

[2] R. Blagus, L. Lusa. “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, 2013 – Springer, pp.2-16.

[3] G.A.Sharifai, Z. Zainol. “Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm," *MDPI Genes*, 2020,pp.-2-26.

[4] J. Li,K. Cheng,S. Wang,F. Morstatter,R.Trevino, J. Tang6 H. Liu. “Feature Selection: A Data Perspective,” ACM Computing Surveys,vol. (50), no. 6, 2018,pp 1–45, DOI:https://doi.org/10.1145/3136625.