

# Table of contents

Volume 1099

2021

◀ Previous issue    Next issue ▶

**International Conference on Applied Scientific Computational Intelligence using Data Science  
(ASCI 2020) 22<sup>nd</sup>-23<sup>rd</sup> December 2020, Jaipur, India**

Accepted papers received: 16 February 2021

Published online: 15 March 2021

Open all abstracts

## Preface

**OPEN ACCESS** 011001

Preface

+ Open abstract     View article     PDF

**OPEN ACCESS** 011002

Peer review declaration

+ Open abstract     View article     PDF

## Papers

**OPEN ACCESS** 012001

Urban Sound Classification Using Convolutional Neural Network Model

Srishti Garg, Tanishq Sehga, Aakriti Jain, Yash Garg, Preeti Nagrath and Rachna Jain

+ Open abstract     View article     PDF

**OPEN ACCESS** 012002

Sentiment Analysis, Tweet Analysis and Visualization on Big Data Using Apache Spark and Hadoop

Sujala D Shetty

+ Open abstract     View article     PDF

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy.

012002 

## Prediction of Presence of Breast Cancer Disease in the Patient using Machine Learning Algorithms and SFS

V Chaurasia, MK Pandey and S Pal

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012004

## A New Tool to Solve Machine Learning Problems under Intuitionistic Fuzzy Sets

R. N. Saraswat and Sapna Gahlot

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012005

## Applications of Flying Ad-hoc Network During COVID-19 Pandemic

Manisha Devi, Sunil Kumar Maakar, Deepak Sinwar, Mahesh Jangid and Poonam Sangwan

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012006

## Computing and Analysis of Multi Server Queueing Model Using Pentagonal Fuzzy Numbers

Mridula Jain and Anamika Jain

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012007

## An Analytical Approach to Predict Employability Status of Students

Bhavna Saini, Ginika Mahajan, Harish Sharma and Ziniya

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012008

## Effect of Dimensionality Reduction on Prediction Accuracy of Effort of Agile Projects Using Principal Component Analysis

Ms. Manju Vyas and Dr. Naveen Hemrajani

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012009

## Performance Study of Snort and Suricata for Intrusion Detection System

Neha V Sharma, Kavita, Gaurav Aggarwal and Saurabh Sharma

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012010

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy.





[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012018

Analysis of Exertion with Relative Grading Mechanism in Academics Using Computational Technique

Nagaraju Dasari, Mahendra Parihar and Mohiddin Shaw Shaik

[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012019

Exploratory and Predictive Analytics of User Preferences from Kaggle LEGO-Toys Datasets Using Spark ML

Pritika Bahad, Preeti Saxena and Raj Kamal

[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012020

SRGM using Testing-Effort Function with Uncertainty in Operating Environment

Ramgopal Dhaka, Bhoopendra Pachauri and Anamika Jain

[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012021

Multi Algorithmic Approach to Galactic Swarm Optimization (MAGSO)

Vasanthakumar Kathirvelu and Venkataraman Muthiah-Nakarajan

[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012022

Methodological Approaches to Data Pre-Processing Formalization for Statistical Analysis

D A Pisareva, E A Pakhomova and O V Rozhkova

[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012023

Social Welfare Maximization in Smart Grid: Review

Gaikwad Sachin Ramnath and R. Harikrishnan

[+ Open abstract](#) [View article](#) [PDF](#)

OPEN ACCESS

012024

A Comparative Study of Laplace Transform and Sumudu Transform on  $\bar{H}$ -Function

Alok Bhargava, Ravi Kumar Jain and Garima Agarwal

[+ Open abstract](#) [View article](#) [PDF](#)

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy.



**OPEN ACCESS**

012025

**Fog Computing in Healthcare: A Review**

Kamini Pareek, Pradeep Kumar Tiwari and Vaibhav Bhatnagar

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012026

**Machine Learning and Prophecy of Behavior: A Breakthrough in Artificial Intelligence**

G Saini, Seema and K Mor

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012027

**The Efficient Resource Scheduling Strategy in Cloud: A Metaheuristic Approach**

Shilpa Maheshwari, Savita Shiwani and Surendra Singh Choudhary

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012028

**A Heuristic Review on Analog Performance and Accomplishment of Activation Functions at RTL Level**

Sudhakar Jyothula

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012029

**On the treatment of zero returns in the estimation of log-GARCH model : Empirical study**

Abdeljalil Settar and Mohammed Badaoui

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012030

**Data Stream Clustering for Big Data Sets: A comparative Analysis**

Ankit Kumar Dubey, Rajendra Gupta and Satanand Mishra

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012031

**Adverse Effects of 5th Generation Mobile Technology on Flora and Fauna: Review Study**

Rajesh Kumar, Rabira Geleta, Amit Pandey and Deepak Sinwar

[+ Open abstract](#) [View article](#) [PDF](#)**OPEN ACCESS**

012032

**A Comparative Analysis of Association Rule Mining Algorithms**

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy



[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012033

### Novel Algorithm for Image Classification Using Cross Deep Learning Technique

Jugnesh Kumar, Pradeep Bedi, S B Goyal, Ashish Shrivastava and Sunil Kumar

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012034

### Analysis of Machine Learning Techniques for Detection System for Web Applications Using Data Mining

Jugnesh Kumar, S B Goyal, Pradeep Bedi, Sunil Kumar and Ashish Shrivastava

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012035

### ICT for Cyber Security in Business

R Hristev and M Veselinova

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012036

### Role of Machine Learning in Sustainable Engineering: A Review

Vaibhav Bhatnagar, Shefali Sharma, Anurag Bhatnagar and Lov Kumar

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012037

### A Design Approach for Identifying, Diagnosing and Controlling Soybean Diseases using CNN Based Computer Vision of the Leaves for Optimizing the Production

Raj Kamal, Sadhna Tiwari, Savita Kolhe and Manojkumar Vilasrao Deshpande

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012038

### A Cost-Efficient Proof-of-Stake-Voting Based Auditable Blockchain e-Voting System

Trishie Sharma, C Rama Krishna and Arshdeep Bahga

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012039

### A Computational Model to Automate the Design of Reinforced Concrete Tee Beam Girder Bridge Using Python Ecosystem

Utkarsh Jain, Mahima Sharma and Charanjeet Singh Tumrate

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see

[our Open Access Cookies policy](#) [View article](#) [PDF](#)



- 
- OPEN ACCESS** 012040  
Fake News Detection Using Machine Learning Approaches  
Z Khanam, B N Alwasel, H Sirafi and M Rashid  
[+](#) Open abstract [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012041  
Fitness for Solving SMCP Using Evolutionary Algorithm  
Neetu Gupta, Ajay Rana and Sumit Gupta  
[+](#) Open abstract [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012042  
Text Summarization Techniques and Applications  
Virender Dehru, Pradeep Kumar Tiwari, Gaurav Aggarwal, Bhavya Joshi and Pawan Kartik  
[+](#) Open abstract [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012043  
A Study on Sentiment Analysis of Mental Illness Using Machine Learning Techniques  
Pradeep Kumar Tiwari, Muskan Sharma, Payal Garg, Tarun Jain, Vivek Kumar Verma and Afzal Hussain  
[+](#) Open abstract [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012044  
A Review on Recent Trends in Secure and Energy Efficient Routing Approaches in Wireless Sensor Networks  
Vivek Sharma and Devershi Pallavi Bhatt  
[+](#) Open abstract [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012045  
Mathematical Analysis of Radiating Viscoelastic Unsteady MHD Fluid Flow through an Absorbent Media between Upstanding Equidistant plates with Joule Heating Impact  
T Mehta, R Mehta and M Kumar  
[+](#) Open abstract [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012046  
Statistical Analysis of Influencing Design Parameters for Two Way Slabs  
Gaurav Sancheti, Hinshal Raitka, Vaibhav Bhatnagar and Jagdish Prasad  
[+](#) Open abstract [View article](#) [PDF](#)

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see [our Privacy and Cookies policy.](#)

## A Review on Swarm Intelligence Techniques in Automated Cryptanalysis of Classical Substitution Cipher

Ashish Jain, Santosh Kumar Vishwakarma, Prakash Chandra Sharma and Nirmal Kumar Gupta

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012048

### Energy Efficient WSN Clustering Using Cuckoo Search

Devershi Pallavi Bhatt, Yogesh Kumar Sharma and Anand Sharma

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012049

### Measuring Accuracy of Stock Price Prediction Using Machine Learning Based Classifiers

Ranjeet Kaur, Dr. Yogesh Kumar Sharma and Devershi Pallavi Bhatt

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012050

### Automatic Detection of COVID 19 Infection Using Deep Learning Models from X-Ray Images

Anju Yadav, Vivek Kumar Verma, Vipin Pal and Saumya Singh

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012051

### Study of Baseline Cyber Security for Various Application Domains

Mudit Chaturvedi, Shilpa Sharma and Gulrej Ahmed

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012052

### Analysis of Design for One-Way Reinforced Concrete Slabs using Machine Learning Models

G Sancheti, H Patil, S Sharma and S Goswami

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012053

### Real Estate Cost Estimation Through Data Mining Techniques

Sandali Khare, Mahendra Kumar Gourisaria, GM Harshvardhan, Subhankar Joardar and Vijander Singh

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012054

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy.





Hiya Luthra, T. Arun Sai Nihith, V. Sri Sai Pravallika, R Raghuram Shree, Ankur Chaurasia and Hina Bansal

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012055

### A Critical Review on Nature Inspired Optimization Algorithms

Vishnu Soni, Abhay Sharma and Vijander Singh

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012056

### A Comparative Study for Efficient Covid-19 Detecting Machine Learning Models on CT Images

Ankur Chaturvedi, Divyansh Mishra, Dr. Vikram Rajpoot, Janvi Gupta and Aditi Sharma

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012057

### Retinal Microaneurysm Detection by CNN

R Deepa and N K Narayanan

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012058

### Analyzing Evolution of the Olympics by Exploratory Data Analysis using R

Rahul Pradhan, Kartik Agrawal and Anubhav Nag

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012059

### Prediction and Classification of Lung Cancer Using Machine Learning Techniques

Pragya Chaturvedi, Anuj Jhamb, Meet Vanani and Varsha Nemade

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012060

### Automatic Prediction of Road Angles using Deep Learning-Based Transfer Learning Models

Sparsh Sharma, Vinit Jhaketiya, Ajay Kaul, Abrar Ahmed Raza, Suhaib Ahmed and Mohd. Naseem

[+ Open abstract](#) [View article](#) [PDF](#)

---

**OPEN ACCESS**

012061

### Problems of Analyzing Socio-Political Content of Internet Resources Based on Neural Network Technologies

[+ Open abstract](#) [View article](#) [PDF](#)

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy



- 
- OPEN ACCESS** 012062  
Digital Transformation Trends and Innovation  
Teresa Guarda, Joel Balseca, Kevin García, Jairon González, Fabian Yagual and Hernán Castillo-Beltran  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012063  
Survey Paper on Visibility Restoration of Underwater Optical Images and Enhancement Techniques  
Khushboo Saxena and Yogesh Kumar Gupta  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012064  
Linear constrained combinatorial optimization on well-described sets  
Oksana Pichugina and Liudmyla Koliechkina  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012065  
Advanced Glaucoma Detection Using Hybrid Approach and Singular Value Decomposition from Fundus Images  
B S Kirar, G Ahmed, S Sharma and D K Agrawal  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012066  
An Empirical Study on Lean Performance Parameters of Manufacturing Sector  
Lokesh Vijayvargy and Srikant Gupta  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012067  
Survey on Smartphone Securities  
Pragya Vaishnav, Manju Kaushik and Linesh Raja  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012068  
Efficient Predictive Modelling for Classification of Coronary Artery Diseases Using Machine Learning Approach  
Savita, Ganga Sharma, Geeta Rani and Vijaypal Singh Dhaka  
[+ Open abstract](#) [View article](#) [PDF](#)

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our Privacy and Cookies policy.



- 
- OPEN ACCESS** 012069  
A Preservation Technology Model for Deteriorating Items with Advertisement Dependent Demand and Partial Trade Credit  
Himanshu Rathore and Rohit Shiv Ashish Sharma  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012070  
Using Internet of Things in Hypervisor Monitoring -Challenges and Opportunities  
Sandeep Mathur, Anurag Kaushik and Ajay Rana  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012071  
Opinion Classification of Product Reviews Using Naïve Bayes, Logistic Regression and Sentiwordnet: Challenges and Survey  
A Dadhich and B Thankachan  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012072  
Comparison of Pairwise Similarity Distance Methods for Effective Hashing  
Ş Öztürk  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012073  
Software Defined Networking: A review on Architecture, Security and Applications  
Kuntal Gaur, Pranjal Choudhary, Priya Yadav, Ayush Jain and Pradeep Kumar  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012074  
An Optimization of Bitmap Index Compression Technique in Bulk Data Movement Infrastructure  
Manoj Kumar, Tarun Kumar Gupta and Deepak Umrao Sarwe  
[+ Open abstract](#) [View article](#) [PDF](#)
- 
- OPEN ACCESS** 012075  
**Design and Analysis of an Efficient Multi-Relational Decision Tree Learning Algorithm**  
**Chhatten Singh Yadav, Abhishek Kumar, Ankit Kumar and Pankaj Dadheech**  
[+ Open abstract](#) [View article](#) [PDF](#)

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see [our Privacy and Cookies policy.](#)

## A Secure Image Watermarking Scheme Based on DWT, SVD and Arnold Transform

Lalan Kumar and Kamred Udham Singh

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012077

## Classification of Imbalanced Data: Review of Methods and Applications

Pradeep Kumar, Roheet Bhatnagar, Kuntal Gaur and Anurag Bhatnagar

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012078

## Structuring Rich Agriculture using Pervasive Computing

Anurag Bhatnagar, Nikhar Bhatnagar and Pradeep Kumar

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012079

## Arbitrage-Type Trade Using Correlation Analysis

Bijesh Dhyani, Pushkar Nigam, Ankit Kumar, Abhishek Kumar, K Venkatesan and V D Ambeth Kumar

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012080

## A Cluster Based Classification for Imbalanced Data Using SMOTE

Rajesh Kumar Tripathi, Linesh Raja, Ankit Kumar, Pankaj Dadheech, Abhishek Kumar and M N Nachappa

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012081

## Model of Using New Media Technology in Higher Education Learning

S Bhavana and V Vijayalakshmi

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**

012082

## A Secure and Efficient Data Migration Over Cloud Computing

G. Madhukar Rao, K. Srinivas, Sayyad Samee, K Venkatesh, Pankaj Dadheech, Linesh Raja and Garvit Yagnik

[+ Open abstract](#) [View article](#) [PDF](#)

**OPEN ACCESS**


012083

## Using Improved MPPT Charge Controller Improvement Functioning Efficiency of Power Grid Connected Solar Photovoltaic System

This site uses cookies. By continuing to use this site, you agree to our use of cookies. To find out more, see our Privacy and Cookies policy.



[+ Open abstract](#)

 [View article](#)

 [PDF](#)

## JOURNAL LINKS

---

[Journal home](#)

---

[Journal scope](#)

---

[Information for organizers](#)

---

[Information for authors](#)

---

[Contact us](#)

---

[Reprint services from Curran Associates](#)

This site uses cookies. By continuing to use this site you agree to our use of cookies. To find out more, see our [Privacy and Cookies policy](#).



PAPER • OPEN ACCESS

## A Cluster Based Classification for Imbalanced Data Using SMOTE

To cite this article: Rajesh Kumar Tripathi *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1099** 012080

View the [article online](#) for updates and enhancements.



**240th ECS Meeting** ORLANDO, FL

Orange County Convention Center **Oct 10-14, 2021**



Abstract submission due: April 9

**SUBMIT NOW**

# A Cluster Based Classification for Imbalanced Data Using SMOTE

**Rajesh Kumar Tripathi<sup>1</sup>, Linesh Raja<sup>2</sup>, Ankit Kumar<sup>3</sup>, Pankaj Dadheech<sup>3</sup>,  
Abhishek Kumar<sup>4</sup> and M N Nachappa<sup>4</sup>**

<sup>1</sup>Graphic Era Hill University, Dehradun, Uttarakhand, India

<sup>2</sup>Department of Computer Applications, Manipal University Jaipur, Rajasthan, India

<sup>3</sup>Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan, India

<sup>4</sup>Jain (Deemed to be University), Bangalore, India

E-mail: lineshreja@gmail.com

**Abstract.** There is tremendous upturn in data repositories because of data generation by various organizations like government, cooperates, health caring in large amounts. Large amount of data is being produced, processed, collected, and analysed online. So there comes a requirement to transform this data into valuable information. This process of extracting the knowledge from large amount of data is referred as data mining. The proposed hybrid approach can be checked on different classifiers like Naïve Bayes, Random forest classifier etc. In proposed methodology we find that SMOTE algorithm which used K-nearest neighbour algorithm is limited to some minority class instances for creating synthetic samples, which sometimes leads to over fitting, so an effective oversampling approach can be developed.

## 1. Introduction

The majority of data in the original word are balanced. This happens if the distribution of the target class among different class levels is not equivalent. This classification of data is one of the toughest problems in machine learning and has become quite important recently. This has contributed to the development of most popular machine learning algorithms to maximize total accuracy, which is the percentage of precise predictions of any classifier. This results in a very low sensitivity and high accuracy to the positive class [1][2]. The best approach is therefore not to concentrate on total precision but to optimize the sensitivities of the positive and negative groups separately. To overcome this problem, several methods have been developed: Samples conform to the previous distribution of the minority and the majority the distribution of balanced classes in the training results. The techniques of sampling can be classified according to basic sampling and advanced methods. Primary sampling techniques include random minority class sampling (RSS), random minority class sampling (ROS) and the composite sampling of both. But with random over-samples of minority data, it is possible that certain minority groups are somewhat enhanced, so that the model is trained in this case leads highly to over fitting. In contrast, random under samples across the majority class lead to the loss of certain important information, since random data are deleted from the majority class [3][4].

SMOTE (Synthetic Minority Oversampling Technique) has been suggested by Chawla et al. (2002), as an advanced over-sampling method. It is intended by creating artificial examples within the minority class to enrich the minority class borders rather than replicating the existing examples in order to avoid overlap. By combining the majority of subgroups with oversampling of the minor classes, multiple re-



modelling for the imbalanced data set has been proposed in order to increase the categorization generalization and to prevent a combination of both sampling methods. Absolute knowledge loss or less stress can increase (Estabrooks et al., 2004). Any data set that can be configured using noise/boundary, redundant/atypical examples, leading to problems with data quality and valid descriptions, however, remove those cases that lead to poorly classified jobs. Batista et al. (2004) proposed the use of Tomek link SMOTE to generate a series of synthetic samples for minority classes using SMOTE to over-sample minority classes, and the use of Tomek link for minority class subclasses where it would eliminate noise and border information. The motivation behind this method is similarly similar to the SMOTE To make connections. He also suggested SMOTE to his nearest neighbor. ENN deletes more cases than to make binding, so more detailed data cleaning is expected. The Unilateral Selection (OSS) method for Kubat and Matwin (2018) is a sub-sampling method that applies CNs to a number of class noise and border instance after applying to make binding's tomek bindings. Confined situations can be called 'insecure' since the wrong side of their decision-making boundaries may contain little terms. The goal of CNN is to eradicate several of these far-reaching decisions. Another method was proposed using the combination of SMOTE and OSS. The SMOTE and OSS methods provide a balancing system for data set distribution, so that the classification results are improved in terms of classification performance. The unilateral selection for subsampling follows the CNN1-NN classification rules, which do not provide sufficient subassembly and often lead to over fitting. Also, unilateral selection does not remove the externalities of the dataset by selecting (Prestanto et al., 2018). So, here we propose an approach that incorporates re-modelling techniques (oversampling and subsampling) for the proper balance of the dataset and improves the nature and quality of the training data by eliminating inconsistent events such as noise / limits / duplicates / advanced data sets that lead to the workplace and all classes are correctly categorized [5][6][7][8][9][10].

## 2. Literature Review

The goal of the proposed OSS subsampling technique is to remove some of these problematic (noise / borderline) examples from the majority class as a form of sub modelling. This suggests to them, it will reduce the examples of the majority class and therefore reduce the classification distinction. According to Chawla et al. (2002): This paper reveals that our (minor) minority sampling groups and minority sampling will obtain higher ratings (ROC space) than the subgroup alone. According to Garcia et al. (2010): The problem of unbalanced learning is related to the performance of the data presented in the data presented and the algorithms present in the presence of severe classification bias. Liu and others. (2013): Sub-Sampling is a popular method for tackling class disequilibrium issues that only uses a subset of the majority class and thus is very efficient. The big disadvantage is that many class examples are ignored. Ganganawar (2012): In this paper we present a short overview of existing solutions to proposed problems of class imbalance at either data or algorithms. One common practice to handle imbalanced data problems is artificial recuperating it by over- and/or under-sampling, which some scientists have proved to be well-integrated with class imbalanced data set, modified support for vector machines, set-based minority-class rule methods [11][12].

According to Lopez et al. (2017): The paper offers a detailed overview of the main issues involved with the use of the internal characteristics of such classified data. It would help to develop the existing models: minor inconsistencies, a lack of focus of training results, duplication of groups, recognition of noise-related results, importance of instances of restrictions and sets adjustment, data for training and testing. We study algorithms on data like features including a view on the actions of certain experimental instances and various approaches and praise [13].

Agarwal etc. (2015): The paper suggested the SCUT hybrid sampling approach for balancing the amount of examples of the training in this multi-class national setting. With the production of synthetic examples, our SCUT numerical accepts minority instances and utilizes cluster analysis for the classification of the major samples [14].



Cao and Zhai (2015): The classification of two numbers of unbalanced data in paper was proposed with a hybrid sampling method. SMOTE is used to generate standardized points for the minority groups, and then the subsampling procedure has been used to remove much of the low-grade samples. Thus comparatively balanced data sets are created and the new data set can be addressed by using SVM [15].

Prestanto and others. (2018): This study will explain the imbalance class in the multiclass EDM dataset management method using a combination of SMOTE and OSS. The class SMOTE and OSS method provide a balancing system for data set distribution, so that classification results improve classification performance [16]

### 3. Proposed Work

In this study, the first step is to gather and then split the unbalanced data collection that we want to define into a testing data set and a test dataset. The model is fitted first to a testing data set, an example set that matches the model parameters. The model is conditioned using a supervised learning approach in a training data set. The test data set is, finally, a data set used to evaluate neutrally how the model fits in with the training data set. The educational curriculum is then split into two subcategories, the minority and the majority.

The primary aim of balancing classes is to increase minority class frequency and reduce majority frequency. This is required to have around the same number of cases for all grades. There is also a re-modeling technique for the balancing class, which determines the solution at the data stage. Two methods are available: over- and under-sampling. Where a minority class is defined in the data sample below, an over-sampling technique is used to expand minority class instances. And a subsample is added to all groups represented as a plurality to offset the minority class. The majority class is sub-sampled and the minority class is over-sampled, so two subsets are formed. We merge them to construct a new equilibrium training package [17][18][19].

We used the classifier to train the new balance training set, and we primarily used the test set created to evaluate the performance of the classifier. We will evaluate the proposed work of publicly available datasets from UCI repositories, KEEL repositories, NASA datasets, etc [20][21][22].

### 4. Performance Evaluation

The proposed approach will be analysed based on the performance measures. Following are the metrics that can be considered such as [23]:

**Confusion Matrix:** Confusion, It's considered an error unit. It is a specific table design that allows the performance of an algorithm, normally a controlled learning one, to be visualized (Shown in Table 1).

**Table 1.** Performance Evaluation of Confusion Matrix

Actual	Predicted	
	Negative Class	Positive Class
Positive Class	False Negative (FN)	True Positive (TP)
Negative Class	True Negative (TN)	False Positive (FP)

- a) True Optimistic (TP): Results are positive and are expected to be positive.
- b) False Negative (FN): False negatives (FN) are good, but negative.
- c) True Talks (TN): The conclusion is pessimistic and optimistic.
- d) False Positive (FP): The observation is positive, but negative.

Precision: The accuracy is referred to as the significant fraction of the identified instances. The exact number of true positives in a class for a class function is the total (i.e. the number of properly identified items in a positive class) separated by an overall number of elements marked as positive elements ( i.e. the sum of genuinely positive and false positive elements, which are items incorrectly marked as class objects).

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

Recall: Remembering the relevant instances is called the fraction. In a classification process, recall is classified by a total number of elements currently of the positive class (i.e. the aggregate of true positive and false negatives, subjects not identified as positive but predisposed to belong to the positive class). Recall is classifying as the number of true positives.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

## 5. Result Analysis

Sensitivity is a significant parameter in the evaluation of the imbalanced dataset classifier efficiency. As with specificity, from the total number of cases present we will calculate the number of positive type instances correctly categorized into our data collection. For imbalances in the data set minority class instances we are most interested in the right definition of minority class instances than in the imbalanced data collection.

Via the use of data resampling, we can create the classification pattern on the training dataset, and then measure the sensitivity of how many minority class instances are properly categorized by the model.

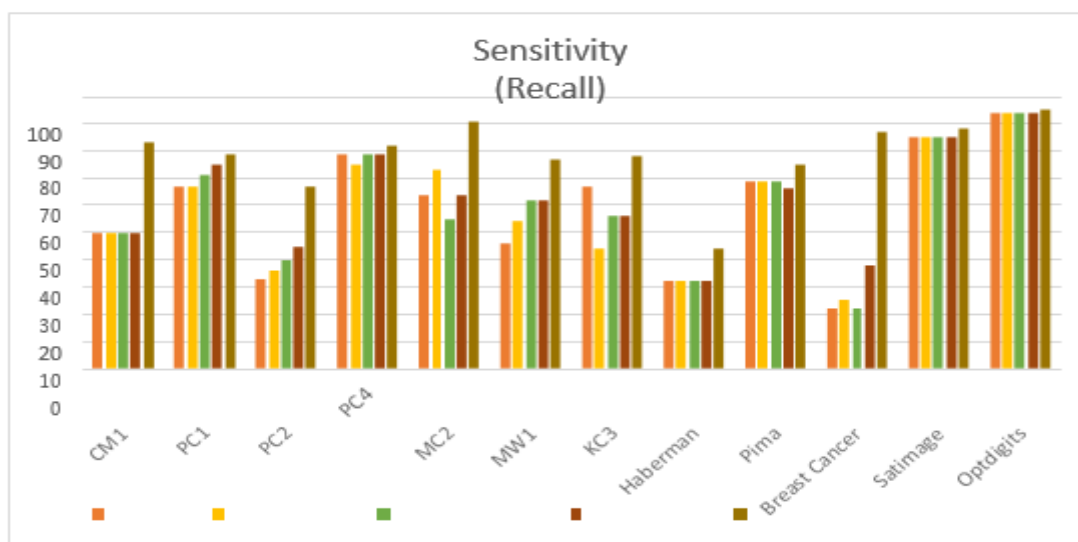
The first column of the table 1 shows the datasets, first row of the table shows data sampling techniques and the entities under are the sensitivity evaluator of the SVM classifier given the sampling technique and the corresponding dataset.

Table 2 and Figure 1 shows the comparison of sensitivity value obtained for all the datasets for existing method (SMOTE, SMOTE-RUS, SMOTE-TOMEK Links and SMOTE-OSS) and proposed method in the form of table and bar graph. From the results, it is observed that the proposed method shown higher sensitivity value as compared to methods SMOTE (Chawla et al., 2002), SMOTE with Random Undersampling (Agrawal et al., 2015), SMOTE with Tomek Links (Batista et al., 2004) and SMOTE with One Sided Selection (Pristyanto et al., 2018). Thus, our proposed method has better performance than existing methods as after removal of all the noisy, borderline, outliers and redundant instances from majority class and oversampling of minority class will lead to the proper classification of instances and will classify minority class instances more properly.

**Table 2.** Performance Comparison in terms of Sensitivity (%)

Dataset	SMOTE (%)	SMOTE-RUS (%)	SMOTE-Tomek Links (%)	SMOTE- OSS (%)	Proposed Hybrid Approach (%)
CM1	50	50	50	50	83
PC1	67	67	71	75	79
PC2	33	36	40	45	67
PC4	79	75	79	79	82
MC2	64	73	55	64	91
MW1	46	54	62	62	77
KC3	67	44	56	56	78
Haberman	32	32	32	32	44

<b>Pima</b>	69	69	69	66	75
<b>Breast Cancer</b>	22	25	22	38	87
<b>Satimage</b>	85	85	85	85	88
<b>Optdigits</b>	94	94	94	94	95



**Figure 1.** Bar Graph of Sensitivity

Sensitivity is an important parameter for evaluating the performance of any classifier. But with class imbalance it is not declared as good parameter because only classifying majority class instances and not correctly classifying minority class instances which are of main interest than also it gives higher accuracy.

## 6. Conclusion

The cluster-based subsampling (CBE) strategy aims to solve class imbalance problems by leaving the majority of instances in the overlapping areas of training information. This was achieved by grouping the training dataset into some  $K$  groups and bypassing all instances of the majority class satisfying  $0 < r < 1$  and  $r$  incomplete / noise instances, unnecessary examples, and the presence of outliers, such that data models have the potential to influence general estimation. For example, sampling techniques such as RSS are another data quality problem. Classification, time of sampling. These problems are inherent in most real-world data sets, as they may be incorrectly created or raised due to the nature of the application domain. It could be argued that these problematic instances can be easily deleted during the data cleansing process, but removing instances blindly from the dataset may worsen the class imbalance problem, depending on the nature of the available information.

## 7. References

- [1] Fernandez, A., Lopez, V., Galar, M., Jesus, M. J. and Herrera, F. 2013. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems* 42: 97–110.
- [2] Ganganwar, V. 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering* 2: 2250-2459.
- [3] García, V., Marqués, A. I. and Sánchez, J. S. 2012. Improving Risk Predictions by Preprocessing Imbalanced Credit Data. In: *Proceedings of 19th International Conference on Neutral Information Processing* held at Doha during November 12-15, 2012, pp. 68-75.

- [4] García, V., Sánchez, J. S. and Mollineda, R. A. 2012. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems* 25: 13–21.
- [5] García, V., Sánchez, J.S., Mollineda, Alejo, R., R., and Sotoca, M. 2007. The class imbalance problem in pattern classification and learning. In: *Proceedings of fourth national conference on data mining and machine learning* held at Spain during September 11-14, 2007, pp. 283-291
- [6] Ghanem, A., Venkatesh, S., and West, G. 2010. Multi-Class Pattern Classification in Imbalanced Data. In: *Proceedings of 20th International Conference on Pattern Recognition* held at Istanbul during August 23-26, 2010, pp. 2881-2884.
- [7] Gray, D., Bowes, D., Davey, N., Sun, Y. and Christianson B. 2011. Further Thoughts on Precision. In: *Proceedings of 15th International Conference on Evaluation and Assessment in Software Engineering* held at Durham during April 11-12, 2011, pp. 129-133.
- [8] Gray, D., Bowes, D., Davey, N., Sun, Y. and Christianson B. 2012. Reflections on the NASA MDP data sets. *IET Software* 6: 549-558.
- [9] Guo, X., Yin, Y., Dong, C., Yang, G. and Zhou, G. 2008. On the Class Imbalance Problem. In: *Proceedings of 2008 Fourth International Conference on Natural Computation* held at Jinan during October 18-20, 2008, pp. 192-201.
- [10] Han, H., Wang, W. Y. and Mao, B. H. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: *Proceedings of International Conference on Intelligent Computing* held at Hefei during August 23-26, 2005, pp. 878-887.
- [11] Hart, P. 1968. The condensed nearest neighbour rule. *IEEE Transactions on Information Theory* 14: 515-516.
- [12] He, H., Bai, Y., Garcia, E. A. and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *Proceedings of IEEE International Joint Conference on Neural Networks* held at Hong Kong during June 1-8, 2008, pp. 1322-1328.
- [13] He, H. and Garcia E. A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21: 1263-1284.
- [14] Holte R. C., Acker L. E. and Porter B. W. 1989. Concept learning and the problem of small disjuncts. In: *Proceedings of the 11th international joint conference on Artificial intelligence* held at San Francisco during August 20 - 25, 1989, pp. 813-818.
- [15] Hsu, C.W., Chang, C. C. and Lin, C. J. 2003. A Practical Guide to Support Vector Classification. *Technical Report submitted to Department of Computer Science*, National Taiwan University Taipei City, Taiwan
- [16] Huang, G., Song, S., Gupta, J.N.D. and Wu, C. 2014. Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE Transactions on Cybernetics* 44: 2405 – 2417.
- [17] Japkowicz, N. 2000. The Class Imbalance Problem: Significance and Strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000)* held at Las Vegas during June 26-29, 2000, pp.111-117.
- [18] Kaur, G. and Singh, L. 2011. Data Mining: An overview. *International Journal of Computer Science and Technology* 2: 336-339
- [19] Pankaj Dadheech, Dinesh Goyal, Sumit Srivastava & C. M. Choudhary, (2018), “An Efficient Approach for Big Data Processing Using Spatial Boolean Queries”, *Journal of Statistics and Management Systems (JSMS)*, 21:4, 583-591.
- [20] Ankit Kumar, Pankaj Dadheech, Vijander Singh, Linesh Raja & Ramesh C. Poonia (2019), “An Enhanced Quantum Key Distribution Protocol for Security Authentication”, *Journal of*

- Discrete Mathematical Sciences and Cryptography*, 22:4, 499-507, DOI: 10.1080/09720529.2019.1637154.
- [21] Ankit Kumar, Pankaj Dadheech, Vijander Singh, Ramesh C. Poonia & Linesh Raja (2019), "An Improved Quantum Key Distribution Protocol for Verification", *Journal of Discrete Mathematical Sciences and Cryptography*, 22:4, 491-498, DOI: 10.1080/09720529.2019.1637153.
- [22] Ankit Kumar, Linesh Raja, Pankaj Dadheech, Manish Bhardwaj (2020), "A Hybrid Cluster Technique for Improving the Efficiency of Colour Image Segmentation", *World Review of Entrepreneurship, Management and Sustainable Development*, Nov. 2020, Vol. 16, Issue 6, pp. 665-679, Print ISSN: 1746-0573 Online ISSN: 1746-0581, <https://doi.org/10.1504/WREMSD.2020.111405>.
- [23] Ankit Kumar, Pankaj Dadheech, Vijander Singh, Linesh Raja (2020), "Performance Modeling for Secure Migration Processes of Legacy Systems to the Cloud Computing", In: *Tin TheinThwel, G. R. Sinha (eds), "Data Deduplication Approaches: Concepts, Strategies and Challenges"*, Chapter-13, pp. 255~280, ISBN: 978-0-12-823395-5, DOI: <https://doi.org/10.1016/B978-0-12-823395-5.00003-3>, Publisher Elsevier.