ISBN Number: 978-93-95305-78-5

DATA IMBALANCE HANDLING TECHNIQUES IN DISEASE PREDICTION MODELS

Deepak Yashwantrao Bhadane
Research Scholar,
P.K. University, Shivpuri, M .P. India, Pin- 473665,
dybhadane@gmail.com
Dr. Indrabhan Supdu Borse
Research Guide,
Department of Computer Science & Engineering
P. K. University, Shivpuri, M .P. India, Pin- 473665, Madhya Pradesh, India
indrabhan2000@gmail.com

Abstract

A class imbalance occurs when there is a significant difference among the two categories of the target variable, with numerous occurrences of one class and few instances of the other. This issue has grown more common in many domains that use models for forecasting, such as illness forecasting techniques, which use information mining and machine learning to solve issues in the healthcare industry. Since the method attends to acquire additional about the larger class owing to its a large sample and acknowledge fewer of the minority class in response to its reduced quantity, class disparity specifically triggers an algorithm powered by machine learning to incorrectly identify events from the minority group even though it can accomplish high precision. This happens since the method might just categorize every case as the a majority class in an information set in class disparity. Public confidence in the consequences of the choice based on the prejudiced result is also impacted by this problem, in addition to the algorithm's forecast outcomes. If this issue of class imbalance is not resolved, the predictive model may label the minority class samples incorrectly, which might undermine the validity of the model's findings. This article offers a thorough analysis of methods for fixing data imbalance and the difficulties associated with doing so.

Keywords: Undersampling. Oversampling, SMOTE, Bagging, Boosting, Class Imbalance, Data Imbalance.

1. INTRODUCTION

Since a business's or organization's capacity to grow and flourish mostly rests on how effectively it comprehends and utilizes the data it has gathered, data has become more vital in today's society. Every

ISBN Number: 978-93-95305-78-5
company or organization nowadays produces massive volumes of data across a range of areas, such as banking, business, finance, and healthcare. Medical data may be provided by hospitals, physicians, care providers, and insurance organizations [1-6]. Upon obtaining the required medical datasets, the next steps would be to analyze and develop suitable modeling algorithms to extract significant information for probable prediction.

Thanks to the rapid improvements in data collection and storage technology, organizations may now get vast amounts of data. The databases are so large that conventional data analysis techniques and instruments are out of date. Data mining is the process of combining sophisticated algorithms with traditional data analysis methods to manage large volumes of data [1-6]. The exciting opportunities it has offered for finding and analyzing new forms of data as well as assessing old types of data in new ways have made the whole process of converting raw data into useful information conceivable.

A branch of computer science known as machine learning (ML) uses a range of statistical, probable, and optimization methods to help computers "learn" from previous encounters and find difficult connections in large, noisy, or complicated data sets. Solutions in medicine, specifically ones that rely on intricate proteomics and genomic information, are especially well-suited for this capacity. As a result, cancer detection and diagnosis heavily rely on machine learning. Machine learning has just recently been used to predict and prognosticate cancer. This latter strategy is especially interesting since it aligns with the emerging trend of predicted personalized medicine. This work was put together by carefully examining the various data imbalance handling procedures, the kinds of data that were utilized, and how well these techniques performed in terms of unevenly distributed data [1]. Due to prejudice against the majority class, imbalanced datasets are more likely to provide altered results, which lower the effectiveness of tools for predictive analysis. This article provides a comprehensive overview of how to deal with unbalanced datasets.

2. PREDICITVE MODELS IN HEALTHCARE

Utilising data, statistics, and machine learning algorithms, predictive modelling is a potent and developing subject in healthcare that aims to enhance patient care, expedite medical procedures, and predict future health outcomes. [1] Thanks to technological developments, the availability of large healthcare datasets, and a growing understanding of its potential to completely transform healthcare delivery, it has been increasingly well-known in recent years. Clinical expertise, data analysis, and

ISBN Number: 978-93-95305-78-5
prediction algorithms are all used in this interdisciplinary approach to help healthcare workers make better decisions, use resources more effectively, and improve patient outcomes while cutting costs.

Fundamentally, predictive modeling in the healthcare industry depends on the extraction and examination of big datasets that include a wide range of patient records, medical histories, test results, and other characteristics relevant to healthcare [1-6]. To produce a thorough information repository, these data sources are painstakingly selected, cleansed, and organised. The development of predictive models that can predict a range of health-related events, including illness diagnoses, treatment outcomes, readmissions, and even the probability that a person would develop a certain medical condition, is based on this resource.

Predictive modeling's capacity to support early illness identification and prevention is among its most important benefits in the healthcare industry [7-9]. These algorithms can identify individuals who might be more susceptible to specific illnesses by examining past data and finding trends. For example, depending on lifestyle choices, genetic predisposition, and past medical data, a predictive model can assist in identifying those who are more likely to acquire diabetes. In order to stop the illness from developing, healthcare professionals can then take proactive measures with tailored treatments, such lifestyle changes or early screenings.

Predictive modeling [1] is also essential for allocating healthcare resources as efficiently as possible. Effectively managing resources, such as beds, personnel, and medical supplies, are a continuous issue for hospitals and healthcare systems. Hospitals may more effectively manage resources and shorten wait times by using predictive algorithms to estimate patient admission rates. In addition to increasing patient pleasure, this guarantees more economical operation of healthcare institutions.

Improving drug adherence is a significant use case for predictive modelling. A common problem in healthcare is prescription non-adherence, which can result in lower health outcomes and higher medical expenses [3-5]. By taking into account a number of variables, such as socioeconomic status, prior adherence history, and the complexity of their drug regimens, predictive models can assist in identifying patients who are at risk of non-adherence. To increase adherence rates and optimise patient outcomes, healthcare practitioners can then use focused interventions, including patient education programs or drug reminders.

Another crucial component of predictive modeling in the medical field is risk stratification. Patients with chronic illnesses, the elderly, or those recuperating from major operations are examples of high-risk

ISBN Number: 978-93-95305-78-5

patients who these models can identify and may need extensive care management. Healthcare providers may lower hospital readmissions, ER visits, and total healthcare expenses by giving high-risk patients priority treatment.

Furthermore, public health programs heavily rely on predictive modelling. Models can forecast disease outbreaks, monitor the spread of infectious illnesses, and evaluate the effectiveness of immunization campaigns by using population-level data. Predictive modelling was crucial in directing public health policy, resource allocation, and vaccination distribution plans during the COVID-19 pandemic, underscoring its vital significance in preserving public health.

In the healthcare industry, patient privacy and ethical issues are crucial when using predictive modeling. Two critical issues that the discipline has to solve are safeguarding private patient information and making sure that forecasts are produced impartially and without prejudice. An continuing ethical conundrum that calls for careful rules and regulations is finding a balance between using data to better medical care and protecting patient rights and anonymity.

One driver that is revolutionizing the healthcare sector is predictive modelling. The way healthcare is provided is changing as a result of its capacity to use data and advanced analytics to improve patient care, predict health outcomes, and allocate resources effectively. Predictive modeling's potential to improve patient outcomes and healthcare systems overall will only increase as technology develops and more data becomes accessible. As predictive models continue to develop and influence the direction of healthcare, it is crucial to approach this subject with ethical considerations at the forefront to protect patient privacy and equity.

3. DISEASE IDENTIFICATION AND RISK ASSESSMENT

Risk assessment and disease diagnosis have become essential components of contemporary healthcare, transforming our knowledge of, approach to, and management of diseases. From conventional clinical evaluations to state-of-the-art artificial intelligence algorithms, these domains cover a wide range of scientific and technical developments with the goal of enhancing patient outcomes and lowering the burden of illness [1-6]. This section explores the complexities of risk assessment and illness diagnosis, emphasizing their importance in modern medicine.

The process of determining a particular sickness or condition in a person based on a collection of clinical signs, symptoms, and laboratory testing is known as disease diagnosis. This essential component of healthcare enables doctors to determine the kind and severity of a patient's condition, allowing them

ISBN Number: 978-93-95305-78-5
to create individualized treatment regimens. From basic physical examinations to advanced imaging procedures like computed tomography (CT) scans and magnetic resonance imaging (MRI), diagnostic methods have changed dramatically over time. These cutting-edge instruments make it easier to identify anomalies and offer comprehensive insights into the interior workings of the body [1-6].

A new age of illness diagnosis has emerged in recent decades as a result of the merging of genetics and molecular biology. For example, genetic testing makes it possible to identify particular genetic variants linked to hereditary disorders, facilitating family planning and early intervention. Furthermore, by customizing therapies to a patient's genetic composition, precision medicine has improved therapeutic results and reduced side effects. In order to diagnose and track conditions including cancer, diabetes, and cardiovascular illnesses, biomarkers—such as protein or gene expression profiles—have become more popular. These biomarkers provide important information about how a disease develops and how well a therapy works.

Risk prediction [1-6] aims to estimate the probability of acquiring specific diseases or health issues in the future, whereas disease diagnostics concentrates on recognising current health concerns. In preventive medicine, risk prediction is essential because it enables medical professionals to take proactive steps to lower the likelihood of disease onset. Cardiovascular risk assessment, which estimates a person's probability of having a heart attack or stroke by taking into consideration variables including age, gender, family history, smoking status, blood pressure, and cholesterol levels, is a perfect illustration of risk prediction. In order to reduce the risk, this information helps doctors prescribe drugs or lifestyle changes.

Artificial intelligence (AI) and data-driven methods have become increasingly potent instruments for predicting illness risk in recent years [8]. To develop predictive models, machine learning algorithms may examine enormous datasets that include details on patient demographics, genetic markers, environmental influences, and medical history. These algorithms have the ability to spot underlying connections and patterns that human therapists would miss. AI-based models, for instance, have been used to forecast the risk of Alzheimer's disease, diabetes, and cancer with very high accuracy. Through the facilitation of early interventions and customised risk reduction measures, this technology has the potential to improve preventive care.

The usage of electronic health records (EHRs) is a prominent illustration of AI-driven risk prediction. EHRs allow healthcare organizations to gather a multitude of patient data, such as clinical notes, test

ISBN Number: 978-93-95305-78-5

results, and medical history [11]. This data may be analysed by machine learning algorithms to find people who are at risk for certain illnesses. An algorithm may, for example, identify trends in EHR data that point to a patient's elevated risk of sepsis, enabling rapid treatment and perhaps saving lives. Additionally, by identifying patients who are at a high risk of hospital readmission, AI-driven prediction models can assist healthcare systems in allocating resources more effectively, hence lowering healthcare expenditures.

The use of mobile health applications and wearable technology in healthcare is another aspect of risk prediction. These devices gather information on a person's heart rate, activity levels, sleep habits, and other variables in real time. Artificial intelligence (AI) systems may evaluate this data to determine a person's risk of obesity, diabetes, or sleep apnoea. Additionally, wearable technology may give consumers tailored feedback and suggestions, enabling them to adopt better lifestyles [13].

Risk assessment and illness diagnosis are essential elements of contemporary healthcare. In the end, they improve patient outcomes and lessen the total cost of disease on society by facilitating early disease identification, individualized treatment, and proactive preventative measures. A new age of precision medicine and preventative care has been ushered in by the merging of cutting-edge technology like artificial intelligence (AI) and genetic testing. The prospect of earlier illness identification and more successful risk reduction grows as we make progress in these areas, portending a better and healthier future for people and communities everywhere.

4. RESEARCH MOTIVATION

In the healthcare industry, predictive data analytics methods have been shown to be very helpful for early disease detection and for enhancing medication and therapeutic treatment, which helps to reduce the rising death rate from diseases like COVID-19, diabetes, heart disease, breast cancer, and chronic kidney disease. By merging mathematical models, computer applications, and clinical data, researchers were able to construct a sizable number of models over the course of several persistent attempts, producing an incredibly valuable multidisciplinary study with societal value. Quality assurance and management are two ongoing activities that oversee quality improvement. The rise in many fatal diseases has been linked to a shift in eating patterns from traditional to fast food, changes in the climate caused by increasing pollution, and a lack of physical exercise.

Early disease detection in this situation will be crucial to keeping the illness under control with the appropriate, closely watched therapy. A preliminary review of some of the already available models

ISBN Number: 978-93-95305-78-5
reveals substantial knowledge gaps in the areas of exploratory data analysis, data treatment analysis, and classifier constraints. By employing hybrid frameworks, one may increase accuracy in terms of synergy with the rapidly expanding mathematical and computational models. Due to prejudice against the majority class, imbalanced datasets are more likely to provide distorted results, which lowers the effectiveness of predictive analytics tools. A thorough analysis of how to deal with unbalanced datasets is suitably provided in view of these findings.

5. DATA IMBALANCE PROBLEM

If the positive and negative values in a dataset are nearly equal, it is said to be balanced. On the other hand, unbalanced datasets are unique situations in which there is an uneven distribution of values across the classes, giving rise to two separate classes: the majority, or negative class, and the minority, or positive class [1, 7, 8, 9, 10]. Both balanced and unbalanced datasets are shown in Figure 1, with the positive values displayed in orange and the negative values in blue.

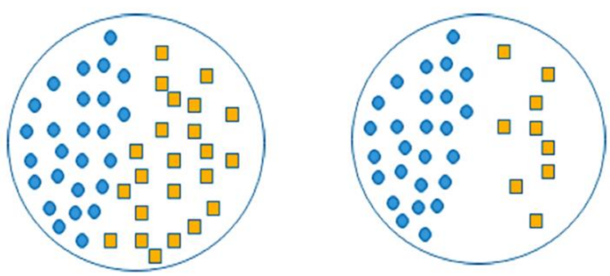

Figure 1. Balanced dataset Vs Imbalanced dataset [1[

6. CLASS IMALANCE PROBLEM TACKLING PROCEDURES

Extremely unbalanced class distributions are a sign of class imbalance issues [87]. A class imbalance may exist in the dataset if samples from one or more classes significantly outnumber those from other classes. Because class imbalance affects the quality and reliability of the output from the machine learning assignment, it has to be addressed using specific techniques and measures. There are three main approaches to addressing the problem of class inequality [1]:

1. Data Driven Methods
2. Algorithm Based Methods
3. Feature Based Methods

ISBN Number: 978-93-95305-78-5

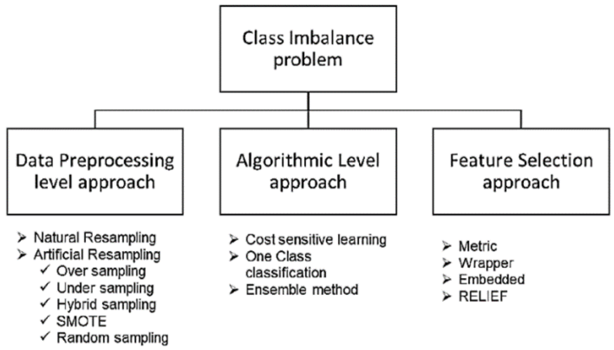

Figure 2. **Different approaches to tackle Class Imbalance Problem [1]**

7. DATA PRE-PROCESSING LEVEL APPROACH

The main objective of this strategy is to resample the data. A pre-processing method called resampling is used to change the class distribution before the model is trained. There are two types of resampling [1, 11, 12, 13, 14, 15]:

7.1 Natural Resampling: The imbalance problem in the dataset is part of its nature of the dataset. the main goal is to the majority class. It's not always possible, but it's simple to obtain.

7.2 Artificial Resampling

It can be accomplished by:

Undersampling: The elimination of cases from the dominant class is carried out at arbitrary from the collection of cases..

Oversampling: instances from the minority class are added to the dataset where the replication is done either randomly or using an algorithm.

SMOTE : S-MOTE oversamples to produce fresh minority class examples by inter-plotting several minority class examples. The freshly generated examples of a majority group are called synthesised observations, and they are obtained by determining the nearest neighbours of each minority group.

Random sampling: The last type of distributed that is produced by the resampling approach can be parameterised to any ratio or fixed to a completely symmetrical allocation. It is the simplest technique to utilise and may be applied to both undersampling and over-sampling.

Random Undersampling: Certain significant and useful information is removed from the dataset in order to compensate for the knowledge loss brought on by under-sampling. For instance, a k-means calculation may be used to implement a cluster-based under-sampling technique.

ISBN Number: 978-93-95305-78-5

Random Oversampling: The minority class's class distribution is introduced to the training dataset after being randomly copied or altered. One way to improve performance is as follows. This approach could be helpful for machine learning algorithms that are affected by a skewed distribution. The outcome is too appropriate.

7.3 Algorithmic methods

Cost-sensitive learning: The learning algorithms charge for inaccurate categorisation. This branch of machine learning focuses on training models using data that has unequal costs or penalties in order to produce predictions. The major objective is to reduce a loss function in order to sway the algorithm in the minority class's favor [16. 17.

One-class classification: The uni-class categorisation branch of machine learning provides techniques for identifying anomalies and outliers. The categorisation target differs from other categories in that it is a one-class objective. It ignores the other classes, the outliers, and only considers the new target class [16, 17]. By fitting a model to the normal data, or OCC for short, one may determine if fresh sample is usual or noise.

Ensemble method: It on each classifier using random subsets during training. Predictions from many classifiers are combined. There are two resample techniques used: bagging and boosting [18, 19, 20].

Bagging approach uses a distinct bootstrap of the dataset to train each classifier. Bootstrap is a subset of N samples that is chosen at random and changed more often. The majority decision of the classifiers is the final result after the models have been trained [1].

Boosting models for classification are trained using the class that is the most difficult to predict. A random sample of the dataset is used to train the first classifier. The majority class in the next sample will be the one with the most records that were incorrectly categorised. On the other hand, although this boosting strategy outperforms bootstrap sampling, it's not necessarily more effective than using just one classifier [21-25.

7.4 Feature Based Techniques

Excessive dataset dimensionality may have a major negative effect on a model's performance. It is thus essential to choose the most beneficial features and get rid of the distracting ones. There are different unique feature based techniques [1, 13, 14, 15, 16]:

ISBN Number: 978-93-95305-78-5

Metrics: It classifies the information by ranking attributes according to their payout.

Wrapper: The algorithm is trained on a selected group of information by ranking features. The framework scores among the most valuable sets of variables based on how well learners learn. When the set of variables presents extremely large, it poses an issue since the parameter selections undergo training multiple times.ability to categorise the documents according to the payout.

Embedded: It seeks to identify the framework's optimal portion of characteristics. An instance is SVM's recurrent features removal method. The relationships between attributes are taken into consideration by integrated and wrapping, although the amount of consideration required significantly rises with large dimensionality datasets [17]. [18]. Even though the metric approach merely counts the role of features rather than the relationships between them, it is the most effective way to handle large data sets.

8. CHALLENGES OF IMBALANCED DATASETS

The following four challenges are involved in dealing with Imbalanced datasets [1, 25-29]

Challenge 1- Bias: The degree of bias between the classes. The closer the ratio between the classes, the higher the chances the classifier will perform better. The dataset can be in different degrees of bias 1:10, 1:00, 1:1000, 1:10000 and so on.

Challenge 2- Overlap: When there is no clear boundary between the two classes, the minority and majority classes occupy the same area. The higher the degree of overlap between classes the more difficult for classifiers to distinguish between them.

Challenge 3- Dataset size: The more samples that artificial intelligence systems have, the greater learning they receive and, thus, the quality of the predictions that is produced.

Challenge 4- Feature Vector size: Whether the data is balanced or not, the classifier has a harder time building an appropriate model the higher the feature vector size. Engineers often do a dimensionality step before applying a machine learning algorithm to the dataset. They just take into account the most important factors in this way. Therefore, if our dataset is also of an imbalanced kind and includes a lot of features, dimensionality reduction may be a really fantastic step to do before you start the research. The present study somewhat addresses the first two problems, bias and overlap. Different sampling techniques have been proposed to reduce bias and the degree of overlap between the majority and minority groups in a dataset.

ISBN Number: 978-93-95305-78-5

9. CONLUSION

Class imbalances have been a difficult issue in the data mining field in recent years. If a particular group in the information has instances that are greater than the others, it is referred to as the dominant class; when there are less instances, it is referred to as the minority class. Because the bigger class had more impacts on the categorization, the circumstance resulted in a less-than-ideal classification. A class imbalance arises whenever any of the groups in the data set has a greater amount of occurrences than the others; these instances are referred to as majority classes and minority classes, respectively. The difficult terrain of uneven data has been thoroughly examined in this book, including everything from its fundamental causes and impacts to a thorough analysis of strategies and evaluation measures meant to alleviate this problem. These principles emphasize the necessity of experimentation, parameter tweaking, and the exploration of different approaches in order to tailor solutions to specific datasets and problem circumstances.

References

1. Govind M. Poddar, Rajendra V. Patill, Satish Kumar N, “Approaches to handle Data Imbalance Problem in Predictive Machine Learning Models: A Comprehensive Review”, *Int J Intell Syst Appl Eng*, vol. 12, no. 21s, pp. 841–856, Mar. 2024.
2. Gaikwad, V. S. ., Shivaji Deore, S. ., Poddar, G. M. ., V. Patil, R. ., Sandeep Hirolikar, D. ., Pravin Borawake, M. ., & Swarnkar, S. K., “Unveiling Market Dynamics through Machine Learning: Strategic Insights and Analysis”,. International Journal of Intelligent Systems and Applications in Engineering, 12(14s), 388–397, 2024.
3. Tarambale, M. ., Naik, K. ., Patil, R. M. ., Patil, R. V. ., Deore, S. S. ., & Bhowmik, M., Detecting Fraudulent Patterns: Real-Time Identification using Machine Learning. International Journal of Intelligent Systems and Applications in Engineering, 12(14s), 2024
4. Iqbal H. Sarrkar, “Machine Learning: Algorithms, Real-World Applications and Research Directions.” SN Computer Science, 2:160, 2021
5. Hafsa Habehh, Sunil Gohel, “Machine in Healthcare”, Current Genomics, vol. 22, issue 4, pp, 291-300, Dec. 2021.
6. Sun, Y., Wong, A. K. C., and Kamel, M. S., “Classification of imbalanced data: a review,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 22, no. 4, pp. 687–719, 2009..

ISBN Number: 978-93-95305-78-5

7. Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, Gong Bing, "Learning from class-imbalanced data: Review of methods and applications" , Expert Systems with Applications,Volume 73, pp. 220-239, 2017.
8. Salim Rezvani, Xizhao Wang, "A broad review on class imbalance learning techniques" , Applied Soft Computing, Volume 143, 2023,\
9. Barandela, R., Sánchez, J. S., García, V., & Rangel, E., "Strategies for Learning in Class Imbalanced Datasets. Pattern Recognition", 36(3), 849-851, 2003
10. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., "Learning from Class-imbalanced Data: Review of Methods and Applications", Expert Systems with Applications", 73, 220-239, 2017.
11. Mukhtar shah, "Imbalanced Data in Machine Learning: A Comprehensive Review", Department of Machine Learning, University of Jumeirah.
12. Chawla, N., et al., "Special issues on learning from imbalanced data sets," ACM SigKDD Explorations Newsletter, vol. 6, no. 1, pp. 1–6, 2004
13. Wongvorachan T, He S, Bulut O., "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining.", Information, 14(1):54., 2023
14. Suwarna J, Kirti N. Mahajan, Yogesh B. Pawar, Yogita D. Bhise, Bharti Jagdale, Rajendra V. Patil, "Plant Growth Analysis using IoT and Reinforcement Learning Techniques for Controlled Environment", Advances in Nonlinear Variational Inequalities, 27(3), pp. 706–715, 2024
15. Sangeeta Kakarwal, Rahul Mapari, Pranoti Prashant Mane, Madhuri Pravin Borawake, Dhanraj R. Dhotre, Rajendra V. Patil, "A Novel Approach for Detection, Segmentation, and Classification of Brain Tumors in MRI Images Using Neural Network and Special C Means Fuzzy Clustering Techniques", Advances in Nonlinear Variational Inequalities, 27(3), pp. 837–872, 2024
16. Akira Tanimoto, So Yamada, Takashi Takenouchi, Masashi Sugiyama, Hisashi Kashima, "Improving imbalanced classification using near-miss instances", Expert Systems with Applications,Volume 201,2022,.
17. Tomek, I., "Two Modifications of CNN", IEEE Transactions on Systems, Man, and Cybernetics (SMC-6): 769-772, 1976

ISBN Number: 978-93-95305-78-5
18. Patil, R. M. ., Patil, R. V. ., Pagare, U. B. ., Navandar, R. K. ., Mapari, R. ., Bhowmik, M. ., & Deore, S. S, "Anomaly Detection in Network Security: Deep Learning for Early Identification.", International Journal of Intelligent Systems and Applications in Engineering, *12*(19s), 133–144, 2024
19. Dennis L Wilson, " Asymptotic properties of nearest neighbor rules using edited data", IEEE Transactions on Systems, Man, and Cybernetics, (3):408–421, 1972.
20. Ajiknya More, " Survey of resampling techniques for improving classification performance in unbalanced datasets", Computer Science Mathematics, arXiv.org, 2016
21. Rajendra V. Patil, Dr. Renu Aggarwal, Dr. Shailesh Shivaji Deore, "Edge Segmentation based Illumination Invariant Feature Detector Phase Congreuncy", 2024 5th International conference on Mobile Computing and Sustainable Informatics (ICMCSI), pp. 91-96, Jan. 2024.
22. K. C. Jondhale , R. V. Patil, "Edge Based technique to estimate Number of clusters in K-means Color Image Segmentation", 3rd IEEE International Conference on Computer Science and Info Tech, Chegndu , China, 2010
23. Ajiknya More, " Survey of resampling techniques for improving classification performance in unbalanced datasets", Computer Science Mathematics, arXiv.org, 2016
24. Smith, M.R., Martinez, T., Giraud-Carrier, C. An instance level analysis of data complexity", Mach Learn 95, 225–256, 2014.
25. Chawla, Nitesh V., et al., "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, no. 1, pp. 321–357, 2002.
26. Rajendra V. Patil, Govind M. Poddar, Bhuvaneshwari Jolad, Bharati Devidas Patil, Dharmesh Dhabliya, Avinash M. Pawar, "Game Theory Applications in Smart Grid Energy Management", Advances in Nonlinear Variational Inequalities, 28(1), pp. 46-60, 2025
27. Sagar V. Joshi, Rajendra V. Patil, Manoj Tarambale, Balkrishna K Patil, Vinit Khetani, Yatin Gandhi, "Stochastic Processes in the Analysis of Electrical Load Forecasting", Advances in Nonlinear Variational Inequalities, 28(1), pp. 15-29, 2025.
28. Nemade, B. ., Bharadi, V. ., Alegavi, S. S., & Marakarkandy, B., " A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons", International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 790–803, 2023

ISBN Number: 978-93-95305-78-5

29. Rajendra V. Patil, Dr, Renu Aggarwal, Ritesh, Sonawane, Govind M. Poddar, Shailesh Shivaji Deore, , Vaishali Turai, “Automatic Marker Generation to Similarity Based Region Merging Algorithm using Edge Information” , Journal of Electrical Systems, 20 (3s), pp. 2227-2240, 2024
30. Bagui, S.S., Mink, D.; Bagui, S.C., Subramaniam, S, “Determining Resampling Ratios Using BSMOTE and SVM-SMOTE for Identifying Rare Attacks in Imbalanced Cybersecurity”, Data. Computers 2023, 12, 204.
31. Rajendra V. Patil., Dr. Renu Aggarwal. Edge Information based Seed Placement Guidance to Single Seeded Region Growing Algorithm. International Journal of Intelligent Systems and Applications in Engineering, 12(12s), 753–759, 2024
32. Rajendra V. Patil, Dr. Renu Aggarwal, Govind M. Poddar, M. Bhowmik. Manohar K. Patil, Embedded Integration Strategy to Image Segmentation Using Canny Edge and K-Means Algorithm. International Journal of Intelligent Systems and Applications in Engineering, 12(13s), 01–08, 2024.