ISSN: 0974-5823 Vol. 7 No. 3 March, 2022

International Journal of Mechanical Engineering

PREPROCESSING OF IMBALANCED ELECTRONIC HEALTHCARE RECORDS USING IMPROVED SMOTE (I–SMOTE) TECHNIQUE

[1]Ms. R. Saranya, [2]Dr. D. Kalaivani.

[1]Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

[2]Associate Professor & Head, Department of Computer Technology, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

Abstract - A well balanced dataset is very important for creating a good prediction model. Medical datasets are often not balanced in their class labels. Most machine learning algorithms work best when the number of samples in each class is about equal. This is because most algorithms are designed to maximize accuracy and reduce errors. However, if the data set in imbalance then In such cases, you get a pretty high accuracy just by predicting the **majority class**, but you fail to capture the **minority class**, which is most often the point of creating the model in the first place. In medical data sets, data are predominately composed of "normal" samples with only a small percentage of "abnormal" ones, leading to the so-called class imbalance problems. In class imbalance problems, inputting all the data into the classifier to build up the learning model will usually lead a learning bias to the majority class. In this study, propose a novel Improved-Synthetic Minority Oversampling Technique (I-SMOTE) which integrates SMOTE oversampling with SMUTE Undersampling. The aim of this paper is to serve as a preliminary study of the potential usefulness of the proposed approach, with the final goal of extending it to utilize local data characteristics. To this end, in the conducted experiments are not only compare the proposed method with the state-of-the-art approaches, but also analyse the factors influencing its performance, with a particular focus on the impact of dataset characteristics.

Keywords: - Class Imbalance, Machine Learning (ML), Medical Data Set, Oversampling, Undersampling, Preprocessing.

1. INTRODUCTION

A balanced dataset is very important for creating a good training set. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. They aim to optimize the overall accuracy without considering the relative distribution of each class. The rapid growth of electronic health records (EHRs) is generating massive health informatics and bioinformatics datasets, and more and more crowdsourced medical data are becoming available [1]. Using statistical data analytics to detect rare but significant healthcare events in these massive unstructured dataset, such as medication errors and disease risk, has the potential to reduce treatment costs, avoid preventable diseases, and improve care quality in general.

One major challenge to effective healthcare data analytics is highly skewed data class distribution, which is referred to as the imbalanced classification problem. An imbalanced classification problem occurs when the classes in a dataset have a highly unequal number of samples. This challenge is often experienced in several disciplines when mining data [2]. The consequence of this bias is that, most classification models developed fail to correctly predict the minority class sample in out-of-sample data. This fact is a huge course of worry for real-world data analysis.

The most frequently used methods to process data are oversampling and Undersampling methods, which balance two classes by increasing minority samples and decreasing majority samples, respectively [3]. Sampling methods based on data are usually simple and intuitive. Undersampling method usually causes information loss while oversampling method tends to balance the original dataset. Thus, the latter one is often adopted in data classification. The proposed work introduces novel algorithm called H-SMOUTE, which is done a critical modification to Synthetic Minority Oversampling Technique (SMOTE) for highly imbalanced datasets, where the generation of new synthetic samples are directed closer to the minority than the majority. In this way, the line of distinction between the two classes will be clearly defined and all samples in data will be located within their class boundaries to ensure accurate prediction of the classifiers developed.

The structure of this paper is organised as follows. In section 2, discusses the overview of related works and also the previous solutions for imbalanced healthcare data are analysed. In section 3, describe the proposed algorithm. The performance evaluation metrics are analysed in section 4. In section 5, concluding remarks and future work are drawn.

2. RELATED WORKS

The problem of learning in imbalance domain has been getting attention in different research areas. The imbalance property that is common to many real healthcare datasets makes classification a challenging task. The imbalanced classification problem in the healthcare domain, where data are often highly skewed due to individual heterogeneity and diversity, affects issues such as cancer diagnostics [9, 10], patient safety informatics [11], and disease risk prediction. In the following, review the strategies and examine their effectiveness when they are combined with standard classifiers to detect medical incidents in imbalanced datasets.

Zhu, M., et al, (2018) deals the classification in class imbalanced data in medical application [4]. Most existing methods are prone to categorize the samples into the majority class, resulting in bias, in particular the insufficient identification of minority class. A kind of novel approach, class weights random forest is introduced to address the problem, by assigning individual weights for each class instead of a single weight. The validation test on UCI data sets demonstrates that for imbalanced medical data, the proposed method enhanced the overall performance of the classifier while producing high accuracy in identifying both majority and minority class.

Gao, L., et al (2020) address the medical image datasets suffer from the imbalance problem [5]. To help address this challenge, one-class classification, which focuses on learning a model using samples from only a single given class, has attracted increasing attention. However, these methods are limited for medical images which usually have complex features. In this paper, a novel method is proposed to enable deep learning models to optimally learn single-class-relevant inherent imaging features by leveraging the concept of imaging complexity. The proposed work investigates and compares the effects of simple but effective perturbing operations applied to images to capture imaging complexity and to enhance feature learning. Extensive experiments are performed on four clinical datasets to show that the proposed method outperforms four state-of-the-art methods.

Xu, Z., et al, (2017) proposed a cluster-based oversampling algorithm (KNSMOTE) combining Synthetic minority oversampling technique (SMOTE) and *k*-means algorithm [6]. However, for imbalanced medical data, the classification accuracy of decision trees-based models is not ideal. The sample classes clustered by *k*-means and the original sample classes are calculated to select the ‘‘safe samples” whose sample classes have not been changed. The ‘‘safe samples” are linearly interpolated to synthesize the new samples. The improved SMOTE sets the oversampling ratio according to the imbalance ratio of the original samples, which is used to synthesize the samples whose number is the same as that of the original samples. The proposed algorithm was applied to the medical datasets, and the average values of the *Sensitivity* and *Specificity* indexes of the Random forest (RF) algorithm were 99.84% and 99.56%, respectively.

Vuttipittayamongkol, P., & Elyan, E. (2020) proposed a framework for predictive diagnostics of diseases with imbalanced records is presented [7]. Early diagnosis of some life-threatening diseases such as cancers and heart is crucial for effective treatments. Supervised machine learning has proved to be a very useful tool to serve this purpose. To reduce the classification bias, propose the usage of an overlap-based Undersampling method to improve the visibility of minority class samples in the region where the two classes overlap. This is achieved by detecting and removing negative class instances from the overlapping region. This will improve class separability in the data space. Experimental results show achievement of high accuracy in the positive class, which is highly preferable in the medical domain, while good trade-offs between sensitivity and specificity were obtained.

Elyan, E., et al (2021) deals the Class-imbalanced datasets are common across several domains such as health, banking, security, and others [8]. The dominance of majority class instances (negative class) often results in biased learning models, and therefore, classifying such datasets requires employing some methods to compact the problem. The proposed work introduces a new hybrid approach aiming at reducing the dominance of the majority class instances using class decomposition and increasing the minority class instances using an oversampling method. Unlike other Undersampling methods, which suffer data loss, our method preserves the majority class instances, yet significantly reduces its dominance, resulting in a more balanced dataset and hence improving the results.

3. PREPROCESSING OF HER USING IMPROVED SMOTE (I-SMOTE)

Generally, many of the Machine learning algorithms applied to classification problems assume that the classes are well balanced. Data sampling is the most common method used to solve the class imbalance problem. The data sampling method involves creating a balanced dataset by adjusting the number of samples of the majority class, which occupies most of an imbalanced dataset, and the minority class, which occupies a small part. The sampling method can be classified as an Undersampling or oversampling method depending on for which of the two classes the number of samples is adjusted [12]. The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling process achieved using additional synthetic data. According to, the

original data obtained using SMOTE is used to synthesize new minority data that are different from the original ones, thereby alleviating the impact of overfitting on the minority class.

The SMOTE is based on the idea of the nearest neighbor algorithm (kNN) and assumes that a synthetic data sample can be interpolated between an original and one of the closest neighbors. The SMOTE algorithm calculates the neighbor environment of each data sample from the minority class randomly selects one of its neighbors and makes synthetic data through the interpolation of data between each sample and the nearest neighbor selected. When the number of synthetic data samples to be made is smaller than the size of the original dataset, the algorithm is randomly selected and an original data sample is used to create synthetic data samples. Conversely, when the number of synthetic data samples to be made is greater than the size of the original dataset, the algorithm iteratively creates synthetic data samples using predetermined oversampling ratio. The SMOTE algorithm is described in detail below:

- Find the k-nearest neighbors for each sample.
- Select samples randomly from a k-nearest neighbor.
- Find the new samples = original samples + difference * gap (0, 1).
- Add new samples to the minority. Finally, a new dataset is created.

The SMOTE method comes with some weaknesses related to its insensitive oversampling where the creation of minority samples fails to account for the distribution of sample from the majority class. This may lead to the generation of unnecessary minority samples around the positive examples that can further exacerbate the problem produced for borderline and noisy in the learning process.

3.1. Improved-Synthetic Minority Oversampling Technique (I-SMOTE)

To perform better prediction, most of the classification algorithms strive to obtain pure samples to learn and make the borderline of each class as definitive as possible. The synthetic examples that are far away from the borderline are easier to classify than the ones close to the borderline, that pose a huge learning challenge for majority of the classifiers. On the basis of these facts, here present a new improved approach (I-SMOTE) for preprocessing of imbalanced training sets, which tries to clearly define the borderline and generate pure synthetic samples from SMOTE generalization. Our proposed method has two stages and discussed as follows:

1. **First stage**, Here, first apply SMOTE algorithm to generate the synthetic instance based on following equation:

$$N = 2 * .r - z/ + z \quad \text{(1)} \qquad \text{… equ. (1)}$$

Where N is the initial synthetic instance number (newly generated), r, is the number of majority class samples, and z, is the number of minority class samples.

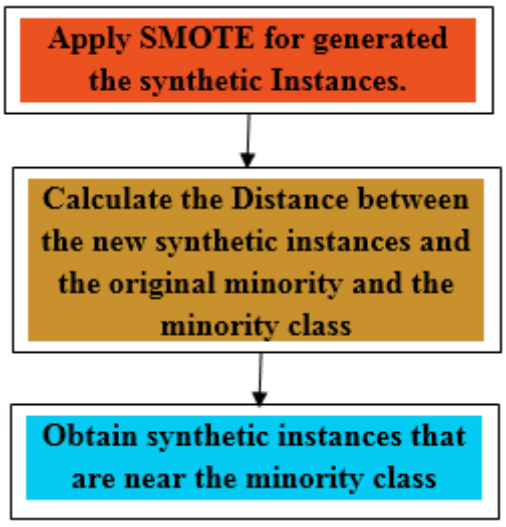

Figure.1: - First stage process of I-SMOTE

2. **Second stage**, To eliminate the synthetic samples with higher proximity to the majority class than the minority as well as the synthetic instances closer to the borderline generated by SMOTE. The A-SMOTE procedure step-by-step is outlined as follows:

- **Step 1:** The synthetic instances that generated by SMOTE might be accepted or rejected on two conditions and it matches with the first stage:
- **Step 2:** After that, with the accepted synthetic instances the following is carried out to eliminate the noisy.
- **Step 3:** Similarly, To calculate the distance.

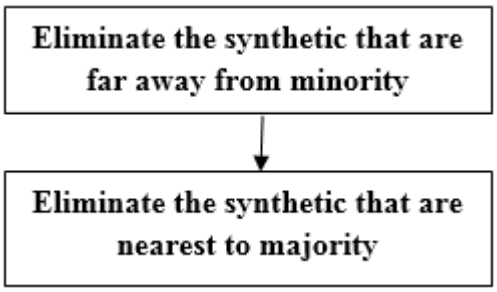

Figure 2: - Second stage process of I-SMOTE

This approach known as I-SMOTE is adopted to design a robust preprocessing method for imbalance learning.

4. EXPERIMENTAL STUDY

In this part, present the experimental design and the results based on the evaluation metrics employed, datasets, different imbalanced methods, and statistical tests. The experiments carried out using MATLAB (2016a). In this research, illustrate the datasets used for the experimental study and the statistical tests used alongside the empirical analysis. The proposed work has used 44 datasets from the KEEL data repository with highly imbalanced rates.

The evaluation criterion is a key factor both in the assessment of the preprocessing performance and guidance of the classifier modeling. In a two-class problem, the electronic health records the results of correctly and incorrectly recognized examples of each class.

4.1. Performance Evaluation Metrics

When preprocessing the image, some researchers are only interested in reducing false-alarm rate, while others may focus on increasing the detection rate. These different requirements can be both satisfied by our metrics.

i. Signal-to-noise ratio (SNR):

The signal to noise ratio (SNR) is a representative of the average signal power to the estimated component present for a pair of original and segmented image. The (SNR) is defined by the equation,

$$SNR = 10 \log 10 \left(\frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (g_{i,j}^2 + f_{i,j}^2)}{\sum_{i=1}^{M} \sum_{j=1}^{N} (g_{i,j}^2 - f_{i,j}^2)} \right)$$

Let $g_{i,j}$ is the original image plus $f_{i,j}$ is the segmented image. i = *1, 2.., M* (range index) and *j = 1, 2.., N* (cross-range index).

ii. Peak Signal to Noise Ratio (PSNR):

The peak signal to noise ratio is the ratio among the highest potential power of a signal and the power of corrupting noise which affects the reliability of its illustration. Since lots of signals have an extremely broad dynamic range, PSNR is typically expressed in terms of the logarithmic decibel scale. The PSNR is most frequently used as a measure of quality of reconstruction in image denoising. This is simply defined by the Mean Square Error (MSE). For 2D M×N monochrome images, the formula for PSNR calculation is given by following equation,

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) = 10\log_{10}\left(\frac{MAX^2}{MSE}\right)$$

Let 'MAX' is the maximum pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255. Higher the PSNR better is the quality.

iii. Mean Squared Error (MSE):

Mean Squared Error (MSE) metric is calculated to measure the change in quality between the original image and filtered image.

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(g_{i,j} - f_{i,j})^2$$

The mean square error is the square of the Euclidean distance among the input and resultant image. In the above equation, $g_{i,j}$ is the original image and $f_{i,j}$ is the estimated image. i = *1, 2.., M* (range index) and *j = 1, 2.., N* (cross-range index).

iv. Structural similarity index (SSIM)

The Structural similarity index (SSIM) is used to quantify structural changes which include luminance, contrast and texture of digital image. It is defined as follow,

$$\frac{\left(2\mu_f\mu_{\hat{f}} + C_1\right)\cdot\left(2\sigma_{f,\hat{f}} + C_2\right)}{\left(\mu_f^2 + \mu_{\hat{f}}^2 + C_1\right)\cdot\left(\sigma_f^2 + \sigma_{\hat{f}}^2 + C_2\right)}$$

From the below results it is found that the proposed filter is outperforming in denoising procedures without losing the useful information such as edges and textures. In Fig 6.2, (a) represents the standard median, the second image (b) represents the DBM, the third (c) represents the WM and the fourth (d) represents proposed method. It is evident from these figures that the above denoised images using our proposed method have better visual quality than that using other filters.

Table 1: Comparison of SNR values

Input Images	Algorithm1	Algorithm2	Algorithm3	Proposed Algorithm
Img1	9.82	9.96	9.83	10.06
Img2	10.47	10.70	10.47	11.08
Img3	11.72	11.99	11.72	12.13
Img4	11.95	12.39	11.96	12.78
Img5	9.77	9.87	9.78	9.93
Img6	7.13	8.14	9.96	13.26
Img7	11.26	12.13	14.78	15.89
Img8	13.95	15.90	16.56	18.78
Img9	13.71	14.69	16.50	17.45
Img10	8.97	9.05	11.90	12.89

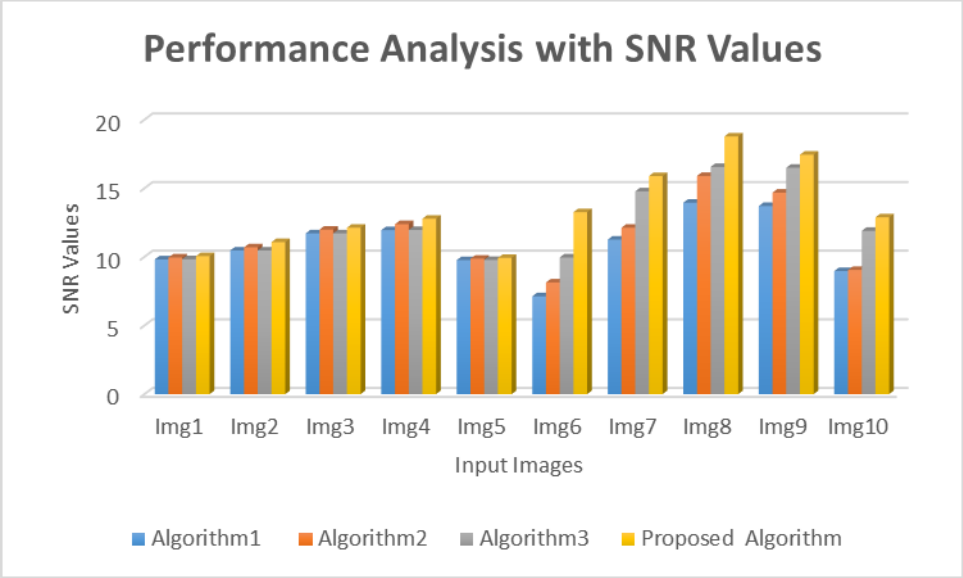

Fig.3. Performance Comparison with SNR values

Table 2: Comparison of PSNR values

Input Images	Algorithm1	Algorithm2	Algorithm3	Proposed Algorithm
Im1	22.45	22.46	24.9	25.1
Im2	23.56	23.07	24.72	25.67
Im3	26.0	29.12	31.6	33.7
Im4	33.00	33.56	34.67	35.72
Im5	39.74	39.98	41.02	42.89
Im6	32.54	32.64	34.1	35.9
Im7	46.0	49.92	51.9	53.1
Im8	39.19	39.67	40.15	42.63
Im9	49.64	49.89	51.20	52.98
Im10	49.09	49.78	50.26	52.74

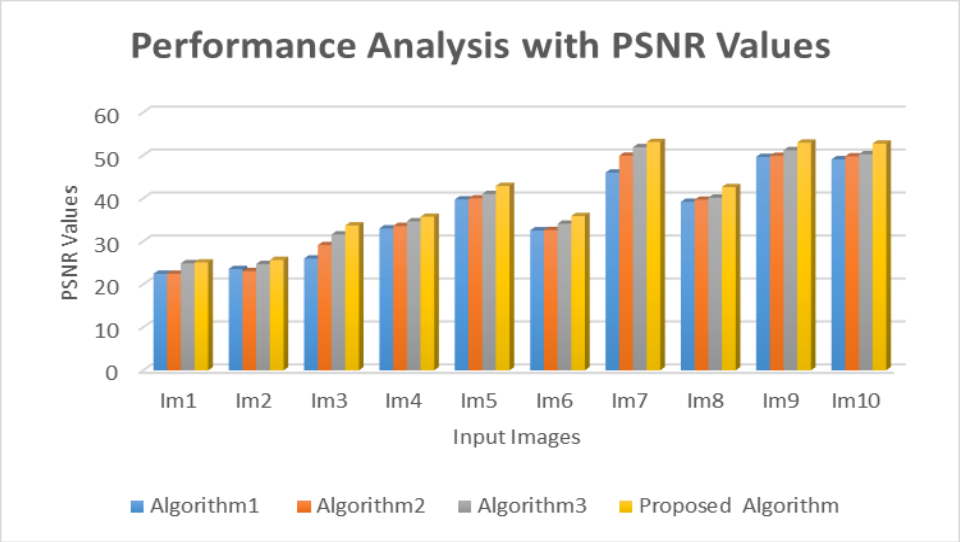

Fig.4. Performance Comparison with PSNR values

Table 3: Comparison of MSE values

Input Images	**Algorithm1**	**Algorithm2**	**Algorithm3**	**Proposed Algorithm**
Im1	3.79	3.71	3.68	3.56
Im2	5.45	5.41	5.37	5.32
Im3	9.07	8.97	9.07	8.45
Im4	8.66	8.52	8.65	8.07
Im5	3.94	3.83	3.93	3.75
Im6	5.19	5.17	2.11	2.02
Im7	5.94	5.56	5.92	5.16
Im8	6.50	6.09	6.49	5.95
Im9	6.00	5.55	4.98	4.05
Im10	3.66	3.89	3.57	3.25

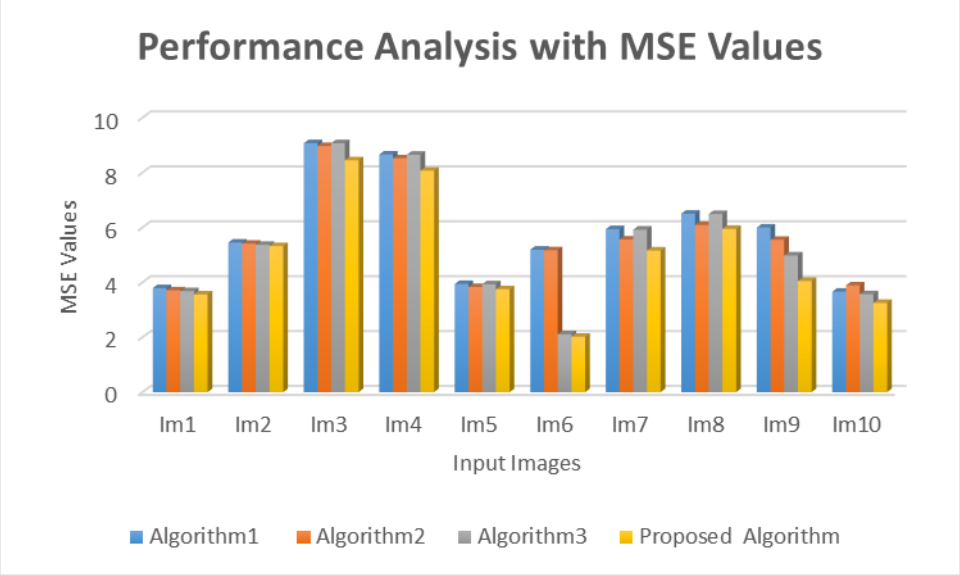

Fig.5. Performance Comparison with MSE values

Table 4: Comparison of SSIM values

Input Images	Algorithm1	Algorithm2	Algorithm3	Proposed Algorithm
Im1	0.91	0.93	0.91	0.95
Im2	0.89	0.91	0.89	0.94
Im3	0.93	0.95	0.93	0.96
Im4	0.94	0.95	0.94	0.98
Im5	0.93	0.94	0.93	0.95
Im6	0.98	0.98	0.99	0.99
Im7	0.97	0.97	0.98	0.98
Im8	0.97	0.97	0.98	0.99
Im9	0.94	0.96	0.96	0.97
Im10	0.92	0.94	0.92	0.96

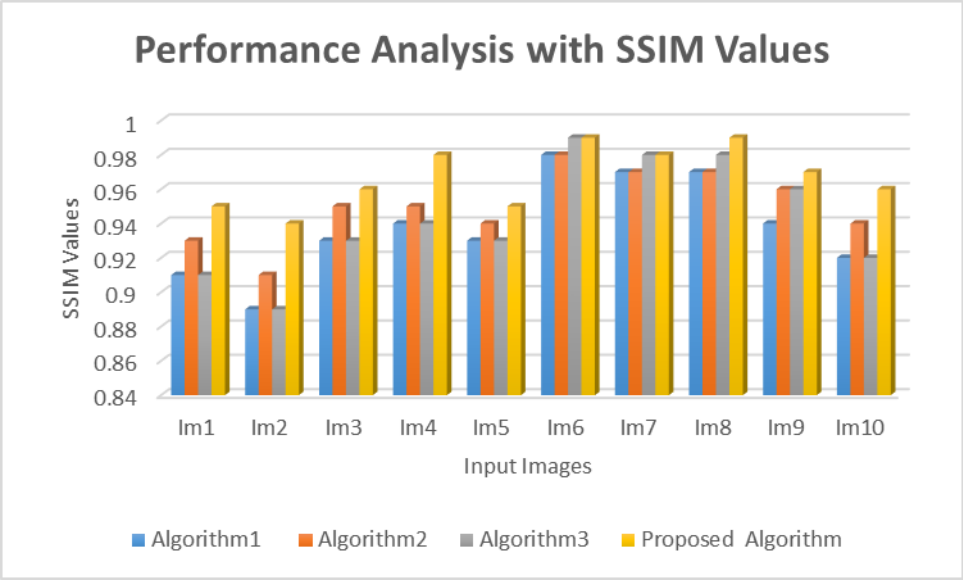

Fig.6. Performance Comparison with MSE values

From Fig.3 – Fig.6, results show that the SNR, PSNR, SSIM gains and MSE decrease to input images for our proposed technique. From the performance evaluation, the proposed method works extremely well for various density of noise is present in the input image.

6. CONCLUSION

Medical data sets usually face class imbalance problems, which will decrease the classification performance because the learning algorithms often overfit the majority class data sets. This research first balances the data size of each class by reducing the data in the majority class and adding virtual samples to the minority one. In this study, proposed a novel approach for highly imbalanced datasets, I-SMOTE, this is an improvement on SMOTE techniques. The proposed I-SMOTE can be a useful tool for researchers and practitioners since it results in the generation of high-quality data. Hence, conclude that this paper presents a useful approach to deal with the class imbalance problem in medical data sets.

REFERENCES

[1] Wosiak, A., & Karbowiak, S. (2017, September). Preprocessing compensation techniques for improved classification of imbalanced medical datasets. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 203-211). IEEE.

[2] Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016). Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS) (pp. 225-228). IEEE.

[3] Hussein, A. S., Li, T., Yohannese, C. W., & Bashir, K. (2019). A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE. International Journal of Computational Intelligence Systems, 12(2), 1412-1422..

[4] Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access, 6, 4641-4652.

[5] Gao, L., Zhang, L., Liu, C., & Wu, S. (2020). Handling imbalanced medical image data: A deep-learning-based one-class classification approach. Artificial Intelligence in Medicine, 108, 101935.

[6] Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N., & Han, X. (2021). A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data. Information Sciences, 572, 574-589.

[7] Vuttipittayamongkol, P., & Elyan, E. (2020, June). Overlap-based undersampling method for classification of imbalanced medical datasets. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 358-369). Springer, Cham.

[8] Elyan, E., Moreno-Garcia, C. F., & Jayne, C. (2021). CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. Neural computing and applications, 33(7), 2839-2851.

[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463–484, 2012.

[10] H. Hassanzadeh, T. Groza, A. Nguyen, and J. Hunter, "Load balancing for imbalanced data sets: classifying scientific artefacts for evidence based medicine," in PRICAI 2014: Trends in Artificial Intelligence. PRICAI 2014, D. N. Pham and S. B. Park, Eds., vol. 8862 of Lecture Notes in Computer Science, pp. 972–984, Springer, Cham, 2014.

[11] Z. S. Y. Wong, "Statistical classification of drug incidents due to look-alike sound-alike mix-ups," Health Informatics Journal, vol. 22, no. 2, pp. 276–292, 2016.

[12] Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2014). Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. *Applied Soft Computing*, *22*, 511-517.