iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54074

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

An Oversampling Algorithm combining SMOTE and RF for Imbalanced Medical Data

Najiya Sigeef[1], Neha Antony[2], Sonu Saji[3], Prof. Shyama R[4]
[1, 2, 3]*Computer Science and Engineering Department, Adi Shankara Institute Of Engineering And Technology, Kalady, Kerala, India*
[5]*Asst Professor, Computer Science and Engineering Department, Adi Shankara Institute Of Engineering And Technology, Kalady, Kerala, India*

Abstract: ***A dataset is said to be imbalanced when the distribution of classes is unbalanced. In medical datasets, this is a typical issue because the positive class, such as the existence of an illness, is frequently rarer than the negative class, such as the absence of sickness. Since machine learning algorithms are prone to being biased in favor of the majority class, this might negatively affect their performance. An oversampling algorithm that combines the SMOTE and KNN SMOTE methods has been presented as a solution to this issue. Synthetic minority oversampling, a well-known oversampling approach, interpolates between existing samples to produce synthetic samples for the minority class.***
The KNN, on the other hand, chooses K's nearest neighbors, mixes them, and produces synthetic samples in space. The oversampling algorithm can be used to balance the dataset, and then the random forest decision tree approach can be used to create a classification model. The random forest algorithm chooses random samples from a set of data and builds a decision tree for each training set of data. Voting is used to determine the decision tree's average, and the prediction result with the highest number of votes is chosen as the final prediction outcome.
Keywords: SMOTE; RF; Imbalanced medical data

I. INTRODUCTION

Medical data is critical to medical research and treatment decisions. However, class imbalance is common in medical datasets, when the number of occurrences in one class is much lower than the other. This might pose issues in machine learning applications by biasing the model towards the majority class and causing it to perform poorly on the minority class.

Various strategies have been proposed to address this issue, including oversampling the minority class using synthetic data generation techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and its derivatives. These strategies can generate fresh samples for the minority class based on existing data, making it easier for the model to learn the minority class's patterns.

In other circumstances, however, oversampling alone may not be sufficient to balance the dataset. In such cases, KNN SMOTE is used to find safe areas for augmentation and fake data points for the minority class are generated, which is then oversampled using SMOTE. KNN SMOTE is a method that has demonstrated good results in balancing unbalanced medical datasets. In this article, we will look at the concept of imbalanced medical data and how KNN SMOTE and SMOTE can help solve it. We will also go over the benefits and drawbacks of these methodologies, as well as practical considerations and iterations for their use in healthcare research. Oversampling's purpose is to balance the dataset by increasing the number of examples from the underrepresented class. This can be accomplished by simply replicating instances from the underrepresented class or by constructing synthetic instances from the underrepresented class. By balancing the dataset, a machine learning model trained on the oversampled dataset should be able to generalize to new, previously unknown data and be less biased towards the more prevalent class.

SMOTE (Synthetic Minority Oversampling Technique) and KNN-based oversampling are strategies for dealing with imbalanced medical data in which one class may be greatly underrepresented in comparison to the other class(es). SMOTE's goal is to generate synthetic examples of the minority class to balance the dataset. It works by picking a minority class instance and looking for its closest neighbors. Then it generates synthetic instances by interpolating between the chosen instance and its nearest neighbors. The synthetic instances that result are added to the original dataset, resulting in an oversampled dataset that is balanced between the minority and majority classes. KNN-based oversampling uses a three-step approach to determine safe areas for augmentation and generate synthetic points for the minority class.

It operates by selecting K's closest neighbors, combining them, and generating synthetic samples in space. The synthetic instances

©IJRASET: All Rights are Reserved

that result are added to the original dataset, resulting in an oversampled dataset that is balanced between the minority and majority classes. Both SMOTE and KNN-based oversampling aim to balance the medical dataset by increasing the number of examples from the underrepresented class. This can help improve the performance of a machine learning model trained on the dataset since it will be less biased towards the more prevalent class and will be better able to generalize to new, previously unknown data.

II. RELATED WORK

When it comes to medical data, the amount of information in minority samples cannot match the amount of information in majority samples. The information in the majority of samples overwhelms the information in the minority samples, resulting in multiple misclassifications. Many advances are being made in this field, and various results are being produced. Among them are the following:

Decision tree classifiers (DTCs) are used in a variety of applications, such as radar signal classification, character identification, remote sensing, medical diagnostics, embedded systems, and speech recognition. DTC evaluates a sample solely against certain groups of classes, eliminating the need for further calculations. [1] FABC is an AdaBoost feature-based classifier that learns from manually labeled photos. The feature vector is a comprehensive collection of measurements at various spatial scales, including filter results and structure likelihood. The Adaboost classifier is less prone to overfitting since the input parameters are not jointly optimized [2]. The multi-class imbalanced problem (MDO) is a technique for oversampling.

It over-samples minority classes by considering each prospective sample and constructing a new synthetic instance with an equal distance from the anticipated class mean. It increases a classifier's generalization capabilities. [3]

Bagging, boosting, and hybrid techniques are being investigated to address two-class imbalanced data sets. It provides a taxonomy for ensemble-based ways to deal with the imbalance problem in classes. It investigates several families of algorithms based on the ensemble learning algorithm on which they are built. [4] To solve unbalanced learning challenges, the new method MWMOTE employs a clustering strategy to generate synthetic samples from weighted, informative minority class samples.

On minority-class samples, it uses both partitioning and over-sampling and outperforms other approaches in terms of accuracy, precision, F-measure,G-mean, and AUC.[5]The entropy-based Support Matrix Machine is a method for analyzing fuzzy membership for unbalanced data sets using entropy. On imbalanced data sets, it may ensure the relevance of the positive class and give more attention to patterns with higher class certainties, creating more flexible decision surfaces than classic SVM, FSVM, and B-FSVM. [6]

RetainVis is a visual analytics system designed to increase the interpretability and interactivity of recurrent neural networks (RNNs) in electronic medical records (EMRs). It blends RetainEX with visualizations to help users explore actual EMRs, gets new perspectives, and formulate new theories. It intends to widen the method's application to a broader range of medical records. [7] The Hybrid Prediction Model (HPM) can provide early prediction of type 2 diabetes and hypertension based on risk factors supplied by people. Three benchmark datasets were used to predict the likelihood of developing diabetes and hypertension. The IoT-based healthcare monitoring system can use the HPM in real-world settings. According to HPM, age, blood pressure, and diabetes were all substantially related. The HPM prediction findings can be accessed via an Android app, allowing users to identify early diabetes and hypertension risks. [8] The goals of creating a MOGP framework for imbalanced data classification, comparing two Pareto-based fitness techniques, and modifying the method to evolve accurate and diverse ensembles were the goals of evolving diverse ensembles using genetic programming for classification with unbalanced data.According to experimental data, MOGP ensembles produced a trustworthy collection of genetic programme classifiers along the trade-off frontier between minority and majority classes. generated by[9] An adaptive synthetic sampling method called ADASYN has been introduced to cope with the two-class classification problem for unbalanced data sets.

It generates more synthetic data for the more difficult-to-learn minority class samples and shifts the classifier decision boundary to focus on those examples. The results of five data sets utilized in simulations show that it is effective. Unbalanced learning is a difficult topic in data mining, machine learning, and AI.[10]

III. PROPOSED METHOD

A prevalent issue in the categorization of medical data is imbalanced datasets, which lead to bias in favor of the dominant class. The classification algorithms might perform poorly as a result of this. An oversampling technique that combines SMOTE and KNN is suggested in order to increase the accuracy of classification results. By extrapolating between close samples, the SMOTE oversampling technique creates synthetic samples in the minority class.

A supervised learning algorithm is KNN. In order to produce the synthetic samples in space, KNN chooses K's nearest neighbors,

combines them, and does so. The oversampling algorithm can be used to balance the dataset, and then the random forest decision tree approach can be used to create a classification model. The random forest algorithm chooses random samples from a set of data and builds a decision tree for each training set of data. Voting is used to determine the decision tree's average, and the prediction result with the highest number of votes is chosen as the final prediction outcome. By choosing representative samples from each group of the original and synthetic samples, KNN lessens the overfitting issue. The generated synthetic samples of the proposed system are more likely to be closer to the real data than other oversampling systems. Additionally, fewer synthetic samples are generated than with conventional oversampling approaches, which reduces the computing time.

IV. SYSTEM ARCHITECTURE

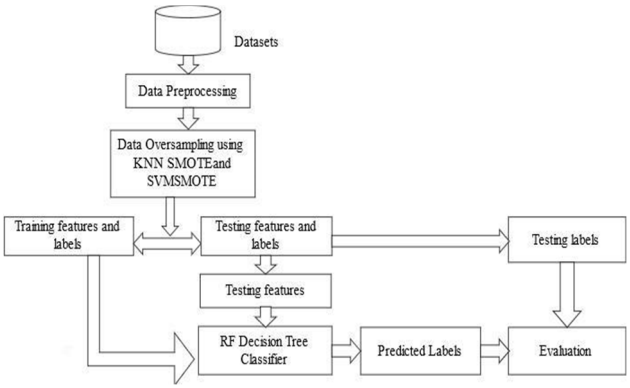


Fig. 1. System Architecture

In order to balance the input data, a number of unbalanced medical datasets must be oversampled. The input data is preprocessed in the data preprocessing module to remove noise, missing values, and pointless characteristics. The data is then split into training and testing sets. To create synthetic samples in the minority class, the SMOTE algorithm is applied to the training set. By extrapolating between close samples, the SMOTE oversampling technique creates synthetic samples in the minority class. A data point is classified using KNN, a supervised learning method, based on the classification of its neighbors. It archives all cases in its database and groups fresh cases according to characteristics in common.The balanced training set is used to train a classification model, such as decision trees, random forests, or support vector machines. The trained model is then used to classify the testing set. Accuracy, precision, and recall are just a few of the performance matrices that are used to assess the classification model's effectiveness.The input data is examined by the data preparation module for errors, missing values, and consistency issues. While missing data are imputed, errors are corrected or deleted. Conflicting values for the same feature are settled by using the majority or mean value. This module normalizes the input data to ensure that the range and scale of each characteristic are comparable. From the preprocessed data, a training set and a testing set are produced. On the training set, the classification model is oversampled, trained, and its performance is evaluated on the testing set. The module assesses the class imbalance in the supplied data to determine its magnitude and the degree of oversampling that is required. The oversampling algorithm combining the SMOTE and KNN algorithms is used to identify the classes in the medical dataset that are imbalanced. SMOTE is an oversampling technique that extrapolates between neighboring samples to produce synthetic samples in the minority class. The classifier divides the medical data into various categories. A random forest decision tree classifier uses a wealth of tried-and-true attributes to make predictions about brand-new data. Finally, the effectiveness of the system is assessed by contrasting it with other classification algorithms. The effectiveness of the model is determined by a variety of metrics, including accuracy, precision, and recall.

V. IMPLEMENTATION

The Synthetic Minority Oversampling Technique, or SMOTE for short, is a better method for dealing with imbalanced data when identifying issues. When observed frequencies are signs that are dispersed throughout the possible values of categorical data, the data is said to be unbalanced. One of the simplest supervised machine learning techniques for classification is the KNN smote algorithm. A data point's classification is based on that of its neighbors. Combining the SMOTE method with KNN entails first classifying a new dataset using the features that are available, and then oversampling the minority class using the SMOTE algorithm.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

As it helps to address the class imbalance while simultaneously keeping the natural structure within the minority class, this can be beneficial in enhancing the performance of classification models trained on skewed medical data. SVM SMOTE is another oversampling approach used in addition to the KNN algorithm. It is specifically made to deal with unbalanced medical data sets, where the majority class (for example, healthy patients) outnumbers the minority class (for example, patients with a specific ailment). Overall, by balancing the class distribution and retaining the local structure of the minority class, the adoption of SVM SMOTE can aid in enhancing the performance of machine learning models on unbalanced medical data sets.

There are three main implementation strategies for systems, and they are as follows:

1) The dataset was oversampled
2) Decision Tree application
3) Forecasting using a balanced dataset Oversampling is a machine learning technique that duplicates examples from the underrepresented class at random in order to equalize the distribution of classes in a dataset.

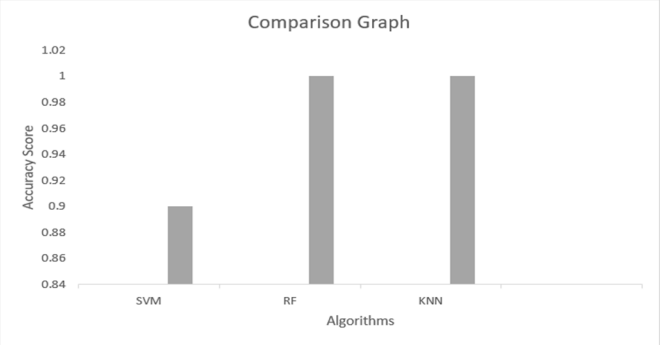


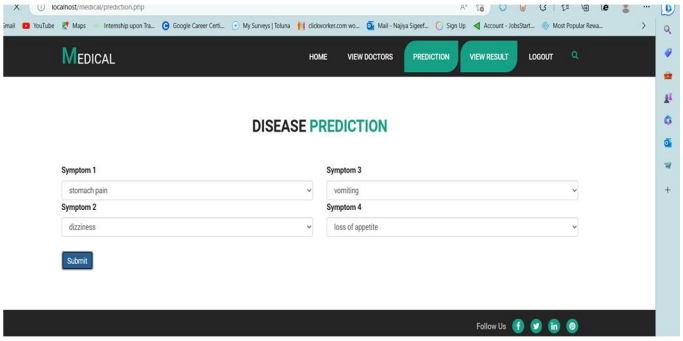
Fig. 2. Principle of SMOTE

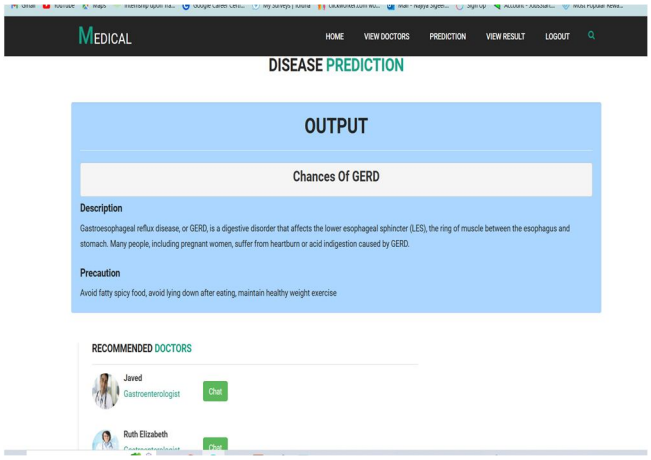
For instance, over-sampling can be used to balance the distribution if a dataset has a class distribution with 90 examples belonging to class B. This can be done by randomly replicating examples from class B until the class distribution is roughly 50/50. ML algorithms such as SVM SMOTE, KNN SMOTE, and others are used to balance the unbalanced dataset. An example of a supervised learning algorithm is a decision tree. Both classification and regression tasks employ it. It functions by building a model of choices that resembles a tree based on the characteristics of the input. By dividing the training data into subsets according to the most important characteristics or features, which are then utilized to make predictions, the model is developed.

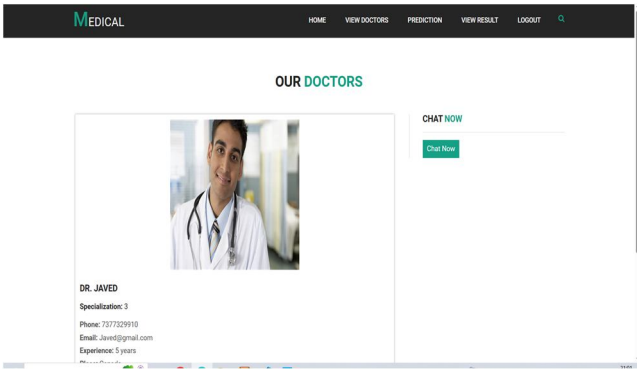
You must carry out the following procedures in order to implement a decision tree:

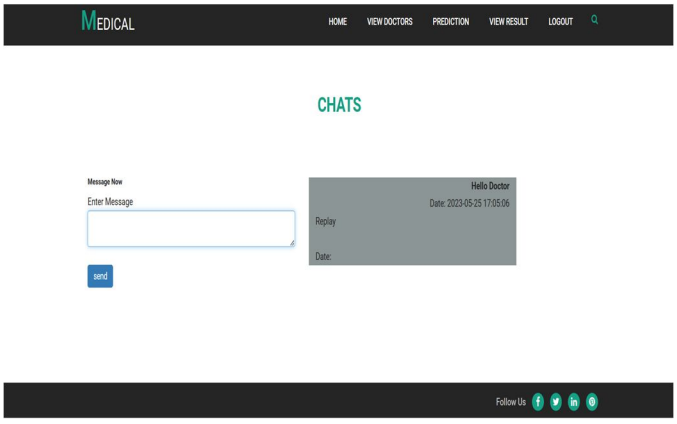
1) *Gather and Prepare the Data:* This involves separating the data into training and test sets, normalizing the data to a common scale, and cleaning the data to handle missing values.
2) *Pick a Measure of Impurity:* This will dictate how the data will be separated at each stage. Entropy and the Gini score are typical indicators of impurity.
3) *Train the Model:* This step entails building the decision tree by iteratively dividing the data into subsets according to the impurity measure of choice. Starting at the root node, the tree is built by selecting the feature that maximizes the impurity measure. Once the leaves are pure, that is, include only examples from one class, the process is repeated for each child node. The random forest Decision tree classifier will be loaded with these tested and trained features.
4) *Generate Predictions:* After the tree has been trained, you can use it to generate predictions based on fresh samples by going from the root node to a leaf node and returning the class that leaf belongs to.
5) *Assess the Model:* The model is then assessed using the testing labels and prediction labels. To gauge how well the model performed, a number of metrics, including recall, accuracy, and precision, can be used. Overall, addressing unbalanced medical data and enhancing machine learning model performance can be accomplished by employing an oversampling technique that combines SVM SMOTE and KNN.

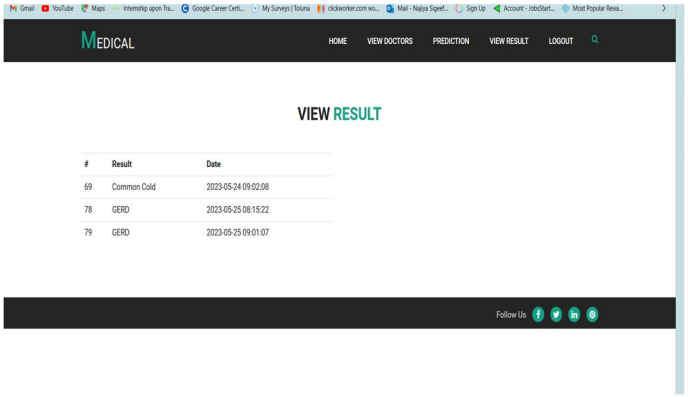
VI. RESULT


MEDICAL
HOME
VIEW DOCTORS
PREDICTION
VIEW RESULT
LOGOUT
DISEASE PREDICTION
Symptom 1
stomach pain
Symptom 3
vomiting
Symptom 2
dizziness
Symptom 4
loss of appetite
Submit
Follow Us


MEDICAL
HOME
VIEW DOCTORS
PREDICTION
VIEW RESULT
LOGOUT
DISEASE PREDICTION
OUTPUT
Chances Of GERD
Description
Gastroesophageal reflux disease, or GERD, is a digestive disorder that affects the lower esophageal sphincter (LES), the ring of muscle between the esophagus and stomach. Many people, including pregnant women, suffer from heartburn or acid indigestion caused by GERD.
Precaution
Avoid fatty spicy food, avoid lying down after eating, maintain healthy weight exercise
RECOMMENDED DOCTORS
Javed
Gastroenterologist
Chat
Ruth Elizabeth


MEDICAL
HOME
VIEW DOCTORS
PREDICTION
VIEW RESULT
LOGOUT
OUR DOCTORS
CHAT NOW
Chat Now
DR. JAVED
Specialization: 3
Phone: 7377329910
Email: Javed@gmail.com
Experience: 5 years


MEDICAL
HOME
VIEW DOCTORS
PREDICTION
VIEW RESULT
LOGOUT
CHATS
Message Now
Enter Message
send
Hello Doctor
Date: 2023-05-25 17:05:06
Replay
Date:
Follow Us



VII. CONCLUSION

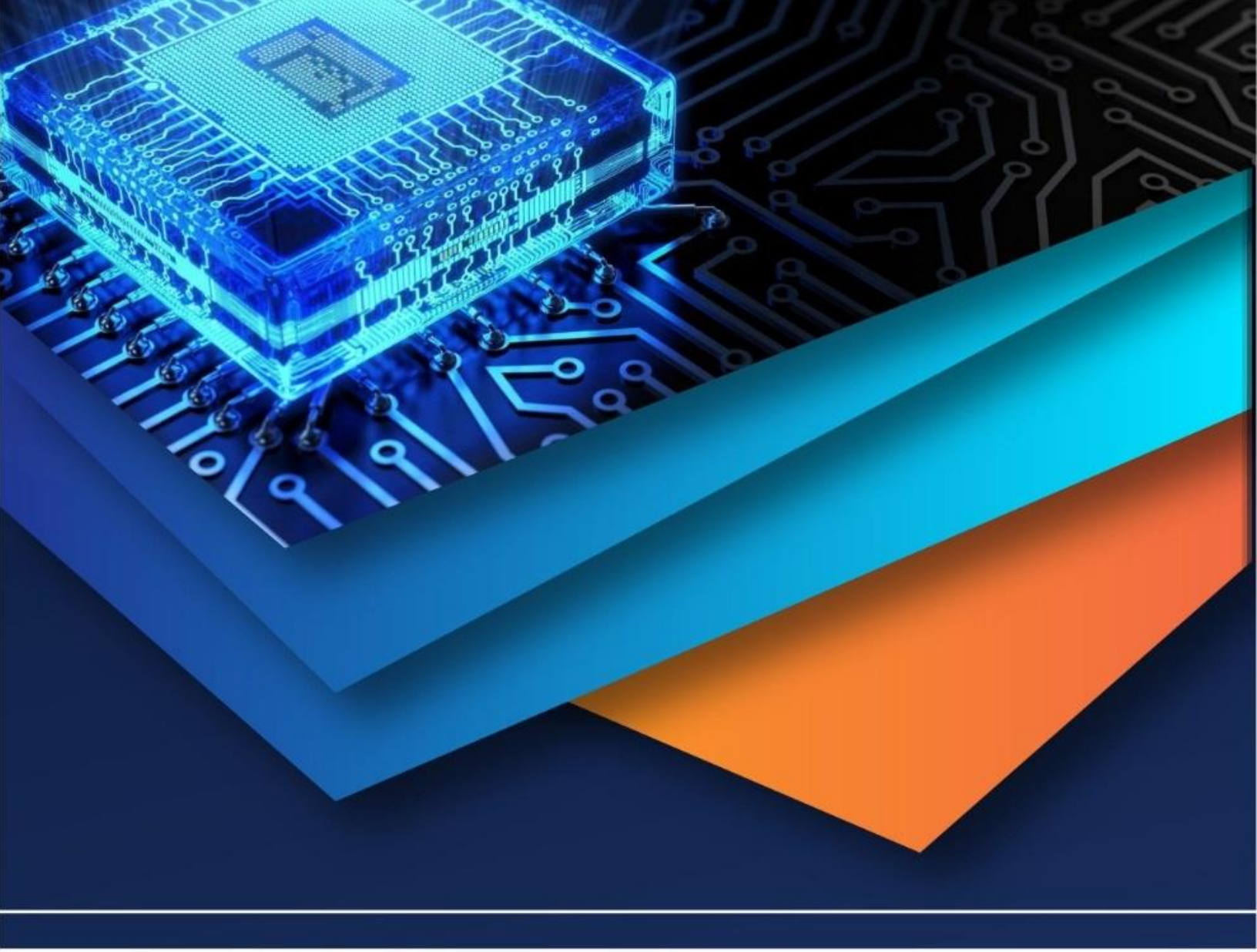
The effectiveness of integrating SMOTE and KNN SMOTE algorithms for skewed medical data is demonstrated by our work. We hope that our work stimulates additional study in this field because we think our method could be a useful tool for managing unbalanced medical data.

VIII. ACKNOWLEDGMENT

We would like to extend our sincere gratitude to The Principal, for providing us with the research facilities and favorable conditions that allowed us to successfully complete the study. The following people and businesses deserve our deepest gratitude for their significant contributions to this effort. First and foremost, we would want to express our gratitude to Prof. Shyama R, our project's mentor, for her leadership and assistance.

REFERENCES

[1] S.R. Safavian, D. Landgrebe. A survey of decision tree classifier methodology[J]. IEEE transactions on systems, man, and cybernetics, 1991, 21(3): 660-674.
[2] Lupascu C A, Tegolo D, Trucco E. FABC: retinal vessel segmentation using AdaBoost[J]. IEEE Transactions on Information Technology in Biomedicine, 2010, 14(5): 1267-1274.
[3] L. Abdi, S. Hashemi. To combat multi-class imbalanced problems by means of oversampling techniques[J]. IEEE transactions on Knowledge and Data Engineering, 2015, 28(1): 238-251.
[4] M. Galar, A. Fernandez, E. Barrenechea, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 42(4): 463-484.
[5] S. Barua, M.M. Islam, X. Yao, et al. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning[J]. IEEE Transactions on knowledge and data engineering, 2012, 26(2): 405-425.
[6] C. Zhu, Z. Wang. Entropy-based matrix learning machine for imbalanced data sets[J]. Pattern Recognition Letters, 2017, 88: 72-80.
[7] B.C. Kwon, M.J. Choi, J.T. Kim, et al. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records[J]. IEEE transactions on visualization and computer graphics, 2018, 25(1): 299-309.
[8] M.F, Ijaz, G. Alfian, M. Syafrudin, et al. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest [J]. Applied Sciences, 2018, 8(8): 1325-1332.
[9] U. Bhowan, M. Johnston, M. Zhang, et al. Evolving diverse ensembles using genetic programming for classification with unbalanced data[J]. IEEE Transactions on Evolutionary Computation, 2012, 17(3): 368-386.
[10] H. He, Y. Bai, E.A. Garcia, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008: 1322-1328.
[11] as: Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, X. Han, An over- sampling algorithm combining SMOTE and k-means for imbalanced medical data, Information Sciences (2021), doi: https://doi.org/10.1016/ j.ins.2021.02.056



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)