Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Improving performance of classifiers for diagnosis of critical diseases to prevent COVID risk

Vinod Kumar\textsuperscript{a}, Gotam Singh Lalotra\textsuperscript{b,*,} Ravi Kant Kumar\textsuperscript{c}

\textsuperscript{a} Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
\textsuperscript{b} Government Degree College Basohli, University of Jammu, India
\textsuperscript{c} Computer Science and Engineering, SRM University, Andhra Pradesh, India

ARTICLE INFO

Keywords: Multi-class classification Clinical dataset COVID-19 Machine learning Class balancing techniques

ABSTRACT

The risk of developing COVID-19 and its variants may be higher in those with pre-existing health conditions such as thyroid disease, Hepatitis C Virus (HCV), breast tissue disease, chronic dermatitis, and other severe infections. Early and precise identification of these disorders is critical. A huge number of patients in nations like India require early and rapid testing as a preventative measure. The problem of imbalance arises from the skewed nature of data in which the instances from majority class are classified correct, while the minority class is unfortunately misclassified by many classifiers. When it comes to human life, this kind of misclassification is unacceptable. To solve the misclassification issue and improve accuracy in such datasets, we applied a variety of data balancing techniques to several machine learning algorithms. The outcomes are encouraging, with a considerable increase in accuracy. As an outcome of these proper diagnoses, we can make plans and take the required actions to stop patients from acquiring serious health issues or viral infections.

1. Introduction

As the number of people infected with Corona Virus Disease (COVID-19) began to rapidly rise, even amongst adults, the entire world was concerned about human life. Because the coronavirus (COVID-19) being a novel virus at the time, it was difficult to determine whether it was more infectious in children, young individuals, or the old, or whether it was equally infectious in all. Scientists later discovered that people who are already sick with a variety of diseases have a higher risk of contracting COVID and experiencing catastrophic complications if they become infected \cite{1,2}. Following this understanding, it has been beneficial to establish plans and decisions to provide work safety rules or immunization plans \cite{3}. As a result, detecting other diseases in people is vital so that they are aware of their present health status and can take the necessary measures, treatments, and immunization regimens. In a country with such a large population as India, keeping proper information of all patients’ histories of critical diseases has become a top priority for the government. Various machine learning algorithms are now accessible for categorization in order to test a big population and assess these massive records of patients’ ailments. The key issue is finding the right dataset to train these machine learning algorithms on. The majority of datasets in the medical area are imbalanced. When machine learning algorithms are trained using this skewed data, the outcomes are skewed. When the amount of data in one class is significantly less than the amount of data in the other

\* Corresponding author.
E-mail address: singh.gotam@gmail.com (G.S. Lalotra).

https://doi.org/10.1016/j.compeleceng.2022.108236
Received 28 December 2021; Received in revised form 3 July 2022; Accepted 13 July 2022
Available online 28 July 2022
0045-7906/© 2022 Elsevier Ltd. All rights reserved.
classes, this is known as class imbalance [4]. As a result, in critical fields, like as medicine, erroneous outcomes due to class imbalanced problems cannot be tolerated. Finally, due to the severity of their current disorders, we are unable to achieve our standards for the appropriate implementation of the plan to protect or save persons from COVID-19. We have included several small variants in our research to solve this key problem and improve the accuracy of machine learning systems. A significant enhancement in the accuracy of machine learning algorithms for the multi-classification of diseases is found, even with imbalanced data.

It is extremely difficult to cure diseases after they have progressed to a critical stage. As a result, obtaining prior information about the sickness is crucial, and machine learning algorithms can be extremely helpful in detecting diseases in less time and with more accuracy. Consequently, we are attempting to improve several machine learning algorithms employing different smote versions in order to address class imbalanced difficulties in the dataset connected to a number of crucial issues.

Hepatitis C is a global disease that affects millions of individuals. The Hepatitis C Virus (HCV), a blood-borne virus that is a major cause of liver cancer, causes hepatitis C, a liver illness [5]. Both acute and chronic hepatitis can be caused by HCV. Chronic HCV infection affects an estimated 71 million persons worldwide. According to the World Health Organization, around 399,000 people died from hepatitis C in 2016 [1]. Antiviral drugs can treat hepatitis C in 95% of instances, however diagnosis and therapy are tough to come by.

Thyroid disease is another frequent condition that affects people all around the world. The gland is a butterfly-shaped organ located in the front and base of the neck [3]. Hormones produced by the thyroid gland play an important role in regulating the functions of various systems throughout the body. Thyroid illness is caused by the thyroid’s overproduction or underproduction of hormones. Thyroid disease is divided into four types: hyperthyroidism, hypothyroidism, thyroiditis, and Hashimoto’s thyroiditis. Infants, teenagers, men, women, and the elderly are all susceptible to this condition [4]. This is a condition that affects people all throughout the world; roughly 20 million Americans have more than one man. Thyroid illness is diagnosed five to eight times more frequently in women than in men, it can be diagnosed by blood tests, imaging tests, or physical exams [5]. In the last two decades, the number of people diagnosed with cancer has nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. The number of cancer-related deaths has also increased, rising from 6.2 million in 2000 to 10 million by 2020. Breast cancer is one of the most lethal cancers in humans, and it is particularly common in women [6].

- Carcinoma (car) is a type of cancer that begins in the cells that make up the skin or the tissue that lines the insides of the organs. Carcinomas are cancerous cells that divide uncontrollably.
- A fibroadenoma (fad) is a benign (noncancerous) tumor. They normally do not hurt and feel like stones in the breast.
- Mastopathy (mas) is a benign, hormone-dependent change in the glandular tissue of the breast.
- Glandular tissue (gla) is the component of the breast that produces milk, and ducts are the tubes that carry milk from the breast to the nipple.
- The breast’s connective (con) tissue gives it shape and support.
- Breast adipose (adi) tissue in the breast is an endocrine system that secretes a variety of growth hormones and enzymes.

Skin illnesses are one of the most common medical issues in the world. Skin disease is a multifaceted notion that encompasses the psychological, social, and financial problems of patients [7]. Prevention is better than cure. Medical datasets are the gold standard for developing machine learning-based illness diagnosis systems that are quick, cheap, and accurate. However, medical datasets of patients with class imbalance problems, which means that a greater number of instances may belong to one class called the majority class and another called the minority class, are considered imbalanced data. The issue of class imbalance persists and is ubiquitous, posing issues in the field of machine learning. Because the overall accuracy and decision-making are biased towards the majority class, any machine learning technique built atop class imbalanced datasets suffers from this imbalance problem, which leads to misclassifying minority class instances or viewing them as noise. Furthermore, the medical dataset may include binary or multi-class labels. Building a machine learning-based efficient diagnosis system for disease prediction is more difficult with a multi-label class imbalanced dataset than with a binary class imbalanced dataset. The categories of the classification process are depicted in Fig. 1. The four datasets-thyroid

Fig. 1. Categories of classification process.
disease, Hepatitis C Virus (HCV), breast tissue disease, and skin disease are multi-class datasets [8]. Therefore, these four multi-class datasets have been taken up and empirically studied by devising a prediction system based on machine learning methods and class balancing approaches. Classification is supervised learning in which target labels or classes are pre-defined and the classification model allocates items in a collection to those labels or classes [9]. For example, deciding whether or not an email is spam. Predicting one of two classes is known as binary classification. A classification task having more than two classifications is referred to as multi-classification. Multi-label classification originated from the investigation of text categorization problems, where each document may be referred to many specified subjects at the same time. Although a little discrepancy is acceptable, most classification datasets do not have the same number of cases in each class.

The presented research effort is one of the solutions for selecting an appropriate model and a class balancing strategy to maximize the model’s performance on multiclass clinical datasets. Diagnostic procedures in the medical industry are both time-consuming and costly. The proposed research output would serve as a tool for categorizing patient disease kinds based on test results and assisting doctors in the clinical decision-making process.

The remainder of the paper is laid out as follows: - Section 2 describes the related works and some important research papers in this field. Section 3 introduces applied data balancing and classification techniques. Section 5 explains the experimental analysis, Section 7 discusses the results and discussion, and Section 8 discusses the conclusion and future work.

2. Related works

Using re-sampling techniques, you may manage a class imbalance in a simple, easy-to-implement, and understandable way. Data is both over- and under-sampled when it is re-sampled. Samples from the majority group are reduced as a result of under-sampling. The process of creating new data from the minority group to match the size of the majority group is known as over-sampling. The complexity of the dataset rises as the number of classes increases [10]. Linear discriminant analysis is a straightforward and effective method for obtaining accurate and timely findings from multi-class data. This study used benchmark datasets, which included a variety of medical datasets [11]. The Random Forest (RF) outperforms the other two Decision Tree techniques used to diagnose heart arrhythmia, namely Classification and Regression Tree (CART) and C4.5, across many databases using 10-fold cross-validation. The MIT-BIH database and the St. Petersburg Institute of Cardiological Technics 12-lead arrhythmia database were used to test the suggested system, which is a combination model of discrete wavelet transformation (DWT), multiscale principal component analysis (MSPCA) de-noising, and RF classifier [12]. Subhani and Anjum combined clinical and gene expression data in their study and used multiple machine learning algorithms on the consolidated dataset. This method, which involves building learning classifiers from large-scale inter domain integrated datasets, produces remarkably accurate clinical predictions. The k-nearest neighbor (k-NN) model has outperformed the other methods on this integrated dataset in terms of accuracy and performance for multi-class classification [13]. To detect one of the sixteen classes, support vector machine (SVM) based approaches such as one versus all, one against one, and error-correcting codes are used to the normalized data [14]. SVM, Neural network (NN), and Logistic Regression (LR) were used to implement over the heart disease dataset in their comparison study, and the performance was measured using F1-score, receiver operating characteristic (ROC) curve, and area under the curve. In their comparison analysis, Khanna et al. [15] concluded that SVM is superior to the other methods under consideration. The UCI repository has been used to pull fifteen real-world medical datasets. Six algorithms are implemented (Naive Bayes, Logistic Regression, K star, Decision Tree: J48, Neural Network: Multilayer Perceptron, and ZeroR), and accuracy is calculated by averaging the results of ten 10-fold cross-validation runs. The Naive Bayes algorithm outperforms the other algorithms on eight of the fifteen datasets [16].

3. Data balancing techniques

Several methods for improving classifiers trained from imbalanced data have already been proposed. The algorithmic level approach and the data level approach are two approaches to the problem of class imbalance. The algorithmic approach tries to modify current classifier learning methods, such as cost-sensitive learning algorithms, to favor the minority group. The data-level strategy aims to rebalance class dispersion by resampling the data space, which involves over-sampling minority class instances and under-sampling majority class instances. Some data-level methods for class balancing operations are listed below. Fig. 9 illustrates the undersampling and oversampling tasks.

| Input: Original Training Dataset |
|----------------------------------|
| Step 1. Pick instance (T_i) from the dataset |
| Step 2. Duplicate minority class examples at random |
| Step 3. while pre-set threshold is met |
| Step 4. Stop |
| Output: Balanced Dataset |

Fig. 2. Pseudo code for oversampling.
3.1. Random oversampling

In random oversampling, samples are picked at random from minority classes and then replaced in the training dataset. This means that in random oversampling, examples from minority classes in the training dataset are duplicated, potentially leading to overfitting of machine learning algorithms. The pseudo-code for random oversampling is shown in Fig. 2.

3.2. Under sampling

The samples are chosen at random and the training datasets are removed. Random under-sampling, on the other hand, eliminates potentially enormous numbers of samples. Defining the decision border between minority instances and majority instances could be difficult. Because of the rejected samples, which lower classification performance. The pseudo-code for random undersampling is shown in Fig. 3.

3.3. SMOTE

A vector space is a collection of feature vectors that represent each sample in an imbalanced dataset with a small number of minority samples compared to a high number of majority samples. For each minority sample $\overline{T_i}$, k nearest neighbors are chosen from the minority sample, and then $\overline{m_i}$ minority sample is chosen at random. Between $\overline{m_i}$ and $(\overline{T_i})$, a point is chosen at random. ($N_{syn}$) is the name of a newly synthesised sample that is contributed to the dataset. The balance parameter $B$ is used to govern the synthesised samples. $B = 1$ denotes that the minority and majority categories each have an equal number of samples. $F$ all refers to the overall number of samples to be synthesised, whereas $F'$ refers to the number of samples to be synthesised from a single minority sample. The minority sample synthesis is repeated $F$ times. The pseudo code for SMOTE [17] is shown in Fig. 4.

3.4. ADASYN

More samples from minority samples are generated along the borderline in ADASYN. The ratio of majority samples in the k nearest neighbors of a minority sample ($t_i$) is $h[i]$. It determines the likelihood of being close to the borderline. It is then normalized in order to calculate $h[i]$ and $F[i]$, the number of samples to synthesize from ($t_i$) [18]. The ADASYN method is implemented from [19].

3.5. SVM SMOTE

The borderline area is determined using this way after the SVM algorithm has been trained. Support vectors were generated from the original training set. Artificial data will be generated at random along the borderline connecting each ‘minority class support vector’ to a few of its nearest neighbors. As a result, it creates an obvious distinction between the minority and majority classes, SVM-SMOTE is implemented from the work the paper [20].

3.6. Borderline SMOTE

Consider the entire training set is S, Q is the minority class ($Q = q_1, q_2, \ldots, q_{num}$) R is the majority class ($R = r_1, r_2, \ldots, r_{num}$) where, $q_{num}$ and $r_{num}$ are the number of minority and majority instances [21].

(1) Start.
(2) In the minority class Q, for each $q_i (i = 1, 2, \ldots, q_{num})$, we compute its g-close neighbors from the whole training set S, the number of majority instances amongst the g-nearest neighbors is represented by $g'(0 \leq g' \leq g)$.
(3) IF $g' = g$, it represents all the k nearest neighbor $q_i$ are majority cases, $q_i$ is noise and it is not used in the subsequent steps. If $\frac{k}{2} \leq g' < g$; the number of $q_i$'s widely held nearest neighbors is higher than the number of its minorities, $q_i$ is easily mis-classified and placed into a set DANGER. If $0 < g' < \frac{k}{2}$, $q_i$ is all right and it is not required to take part in the following stages.

| **Input:** Original Training Dataset |
|-------------------------------------|
| **Step 1.** Pick instance (T_i) from the dataset |
| **Step 2.** Remove members from the majority class at random. |
| **Step 3.** Continue until the pre-determined threshold is reached |
| **Step 4.** Stop |
| **Output:** Balanced Dataset |

Fig. 3. Pseudo code for undersampling.
The data on the margins of the minority class Q are used as examples in DANGER, and we can find that $\text{DANGER} \subseteq Q$. We place $\text{DANGER} = \{q_1', q_2', q_3', \ldots, q_{dnum}'\}$, $0 \leq dnum \leq \text{num}$. From Q, we determine the k closest neighbors of each case in DANGER.

Now, we create $s \times dnum$ artificial positive samples based on the information in DANGER, where s is an integer within $(1, k)$. We choose s nearest neighbor from the k nearest neighbors in Q at random for each $q_i'$. First, we determine the differences, $\text{diff}_{j} (j = 1, 2, \ldots, s)$ between $q_i'$ and its s closest neighbor from, then multiply, $\text{diff}_{j}$ by a random number $r_j (1, 2, \ldots, s)$ between 0 and 1. After all s new artificial marginal cases are produced between $q_i'$ and its closest neighbors:

$$\text{synthetic}_j = q_i' + \eta_j \times \text{diff}_{j}, j = 1, 2, \ldots, s$$

It can be repeated the above process for each $q_i'$ in DANGER and it can reach $s \times dnum$ synthetical examples.

### 3.7. SMOTENC

Below are the steps of the ‘Synthetic Minority Over-sampling Technique-Nominal Continuous [SMOTE-NC]’ [21].

1. **Start.**
2. **Median computation:** The median of standard deviations for all continuous characteristics is estimated for the minority class. If a sample’s minimal features deviate from its likely nearest neighbors, the median is already included in the Euclidean distance. The median is used to adjust the difference in nominal features by an amount proportional to the difference in continuous feature values.
3. **Nearest neighbor computation:** The Euclidean distance between the feature vector for which k-nearest neighbors are being found (in the minority class sample) and the other feature vectors (of minority class samples) is evaluated using the continuous feature space. The median of the standard deviations estimated before, in the Euclidean distance, for each different nominal feature between the considered feature vector and its possible nearest-neighbor.
4. **Populate the synthetic sample:** The new synthetic minority class sample’s continuous features are created with the same SMOTE method as previously shown. The value found in the majority of the k-nearest neighbors is used as the nominal feature.
5. **Stop.**

### 3.8. SMOTEEN

At first, SMOTE specifies the k-nearest Neighbors (k-NNs) (denoted by $\psi_x$) for each minority sample $x_i \in \alpha_{\text{min}}$. For generating a synthetic data model $x_{\text{new}}$ for $x_i$, SMOTE arbitrarily selects an element $\hat{x}_i$ in $\psi_x$ and $\hat{x}_i$ in $\alpha_{\text{min}}$. The feature vector of $x_{\text{new}}$ is the sum of the feature vectors of $x_i$ and the value which is obtained by multiplying ($\hat{x}_i - x_i$) with a random value $\delta$, where $0 < \delta < 1$. In this way, we achieve a synthetic point along the line segment join up $x_i$ and $\hat{x}_i$. Furthermore, the Edited Nearest neighbor (ENN) is utilized to clear up class overlap. The pseudo code for SMOTEEN [22] is shown in Fig. 5.
3.9. SMOTETOMEK

It is a revised version of SMOTE in which the TOMEK links are employed to filter out the noisy data. The TOMEK links are specified as; Instance ‘l’ is the nearest neighbor of instance m where, m is the nearest neighbor of l and l and m belong to different classes [23]. Pseudo code has been mentioned in the Fig. 6.

4. Classification techniques

A concise description of the algorithm of each classification techniques has been described here with the goal of presenting essential insights into six different classifiers.

4.1. Linear Discriminant Analysis (LDA)

The LDA is the preferred linear classification approach when there are more than two classes. The description of LDA is quite simple. It contains statistical data attributes. The mean and variance for each class are determined for a single input variable (x). The LDA model needs the estimation of statistics from the training data for each class.

(1) Mean of each input value.
(2) Probability of an instance.
(3) Covariance for the input data.

LDA model figures out the mean and variance of data for each class and it is easy to compute in case of univariate (single input variable) scenario with two classes. The mean value of each input (x) for each class (k) can be computed as below-

$$\text{mean}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i$$

Where, for the class k, mean\(k\) is the mean value of x, \(n_k\) is the number of instances of class k. The variance is computed through all classes taking the average squared difference of each value from the mean.

$$\sigma^2 = \frac{1}{n-k} \sum_{i=1}^{n_k} (x_i - \text{mean}_k)^2$$

Where, \(\sigma^2\) denotes the variance of all inputs (n), ‘k’ is the number of classes and mean\(k\) is the mean of x for the class to which \(x_i\) belongs, n is the number of instances.

LDA yields estimates by calculating the likelihood that a given set of inputs will fall into each of the classes. The output class with the highest probability stands as the final prediction. The Bayes Theorem is used to calculate the model’s probabilities (Fig. 7).

4.2. Decision Tree (DT)

It is a flowchart-like tree structure in which each internal node represents a test on an attribute, each branch represents the test’s outcome, and class labels have been symbolized by a leaf node, making it a Decision Tree. The root node is the highest node in a tree. It should be acknowledged that no root-to-leaf path should get the same discrete attribute more than once [24] A Decision Tree can quickly create comprehensible rules and do classification tasks with less computation [25]. It can be illustrated simply as in Fig. 8.

| Input: Imbalance Training Data |
|-------------------------------|
| 1. Select \(x_i\) at random from minority classes. |
| 2. Determine \(x_i\)’s k-nearest neighbour ‘ \(\Psi_{x_i}\) ’ |
| 3. Compute \(x_{\text{new}} = x_i + (\bar{x}_j - x_i)\times\delta \) |
| 4. If the balance ratio satisfies if no go to step1 else |
| 5. Employing ENN, remove the noise sample. |
| ENN |
| { For every observation O } |
| Get three closest neighbours of O |
| If O’s three closest neighbours misclassify it, |
| Then delete O |
| End IF |
| End For |
| } |
| 6. End |

Output: Balanced Training Data

Fig. 5. Pseudo code of SMOTEEN.
**Input:** Imbalance Training Data

1. Select $x_i$ randomly from minority classes.
2. Obtain $x_i$'s k-nearest neighbour `ψ $x_i$`
3. Produce $x_{new} = x_i + (\hat{x}_i - x_i)\times\delta$
4. IF balancing ratio is not satisfied go to 1 
   Else
5. Eliminate noise from sample by implementing TOMEK 
   TOMEK $(l, m)$
   `{ 
   m's closest neighbour is l.
   The nearest neighbour of l is m.
   l and m are from distinct classes.
   }`
6. End

**Output:** Balanced Training Data

---

Fig. 6. Pseudo code of SMOTETOMEK.

Fig. 7. Undersampling v/s Oversampling task.

Are you experiencing COVID-19 Symptoms?

- **Yes**
  - Get Tested and if Positive (Diagnostic)
    - **Yes**
      - Contact your Primary care physician
    - **No**
      - Stay home and take care of yourself
  - Get Tested and if positive (Asymptomatic)
    - **Yes**
      - Contact your Primary care
    - **No**
      - Stay home and take care of yourself

- **No**
4.3. Support vector machine (SVM)

V. Vapnik produced a very effective and well-known categorization system in 1992, called SVM. In this the hyperplane mechanism is applied to split two data points that are on opposite sides of the hyperplane. The purpose is to raise the margin so that there is a significant deciding gap between the instances and the hyperplane on both sides. The segregating hyperplane is depicted in Eq. (3).

\[(W \cdot X + b = 0)\]  

Where, \(W = \{w_1, w_2, w_3, \ldots, w_n\}\) designates the weight vector, \(n\): number of attributes, \(X\) is \(n\) dimensional vector; and \(b\) stances for a scalar (a bias).

\[D = \{(x_i, y_i) | x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}\}, \text{Where} \ D \text{ stands for dataset-}\]

(i.) \([W \cdot x_i + b > -1]\) \quad \text{for} \ y_i = 1(\text{Label : class - 1}) \quad (4)

(ii.) \([W \cdot x_i + b < 1]\) \quad \text{for} \ y_i = -1(\text{Label : class 1}) \quad (5)

4.4. Gaussian Naive Bayes (GNB)

If the majority of the attributes in the instances are continuous, Gaussian Naive Bayes is applied. The conditional probability is computed using the equation in Eq. (6). The Bayesian theorem is used for the Naive Bayes classifier technique. When coping with complex classification, the Gaussian Naive Bayes method is preferable.

\[p(X = x | C = c) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} \cdot \frac{(x - \mu)^2}{\sigma^2}\right)\]  

Where \(x = \text{variable}, \ c = \text{Class}, \ \mu = \text{mean and} \ \sigma \text{ is variance of predictor distribution}\)

4.5. k-nearest neighbor (k-NN)

The k-Nearest Neighbor (k-NN) prediction model is widely recognized as a lazy learning (no learning) approach. The prediction is based on the k-nearest numbers supplied to it as input. In general, the Euclidean distance formula is used to calculate the neighborhood. Other distance measures, such as Minkowski, Hamming, and Manhattan distances, are used as needed. The neighborhood distance dist \((x,y)\) from points ‘x’ to ‘y’ is decided using the formula as given in Eq. (7).

\[\text{dist}(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}\]  

4.6. Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) abridge and mimic the activities of brain. An ANN is a set of modules identified as artificial neurons that admit input, modify their inner state (activation) in reply to the input, and produce output based on the activation function and input supplied.

- **Weights (W):** \(\{w_1, w_2, w_3, \ldots, w_n\}\) signifies the neuron strength.
- **Input layer:** this integrates inputs and corresponding weights.
- **Bias (b):** The curve of the activation function is adjusted by Bias.
- **Hidden layer:** The hidden layer has summation as well as activation function. The ANN may be composed of many hidden layers.
- **Activation function:** The activation function, which offers neural networks non-linear features, is a crucial component. It transforms any artificial neuron (AN) input into output. Following that, the acquired output is sent into the next layer of AN. Many activation functions are available such as Radial basis function, SoftMax, Tanh (), etc. The widely used sigmoid function has been mentioned in Eq. (8).

\[\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}\]  

**Output layer:** The set of outcomes generated by the preceding layer is given to Output layer.

5. Experimental design

In this segment, we report a wide-ranging experimental setup that aim to appraise the performance of different multi-classification models with different sets of data balancing techniques as accurately as possible. The whole experiment was carried out by building multi-classification models in “Scikit-learn” a python-based machine learning library. To achieve a class balanced dataset, another python-based python package - imbalanced-learn which offers a number of re-sampling methods normally used in datasets for class balancing. The pictorial and algorithmic representation of the proposed prediction system for multi classification over medical datasets are given in Fig. 9 and Fig. 10, respectively.
Datasets

This research is conducted on four clinical datasets: Thyroid disease, HCV, breast tissue disease, and skin disease and these were downloaded from the UCI machine learning repository. A brief discussion of these datasets is given as follows:

(A) Thyroid disease

The thyroid dataset was composed of 7200 samples comprising 21 attributes and was classified into three different classes. Out of 21 attributes, 15 attributes contain real values and the other 6 attributes have categorical values. The first class represents Hyperthyroid which has 166 instances in it. The second class belongs to Hypothyroid having 368 instances, and the third label is normal (not hypothyroid) and possesses 6666 instances. The research problem is to determine whether a patient referred to the clinic is hypothyroid or not.

(A) Hepatitis C Virus (HCV) for the Egyptian patients

This dataset was collected from Egyptian patients who underwent treatment dosages for HCV for nearly 18 months. It contains 1385 instances with 29 features and four target classes, namely: 'Portal Fibrosis', 'Few Septa', 'Many septa', 'Many septa' with a number of instances of 336, 332, 355, and 362, respectively.

(A) Breast tissue disease

This breast tissue disease dataset was recorded as electrical impedance measurements of newly excised tissue instances from the breast. The dataset contains 106 numbers of instances and consists of 10 features with real values. The dataset can be used for predicting the classification into six classes namely- 'car', 'fad', 'mas', 'gla', 'con', and 'adi'.

(A) Dermatology/ skin disease dataset

This multiclass dataset contains 366 Number of Instances with 33 features of categorical nature and have six target disease classes namely- 'Psoriasis', 'Lichen planus', 'Seborheic dermatitis', 'Pityriasis rosea', 'Cronic dermatitis' and 'Pityriasis rubra pilaris'. The main aim is to categorize the type of Eryhemato-Squamous (i.e., common skin disease) from this dataset.

6. Performance evaluation criteria

For classification purposes, there exist various supervised machine learning models. But for a particular problem, which model is
appropriate, it can only be known by the comparative assessments. Some of the most popular and widely used performance evaluation metrics for classifiers are as follows:

**Contingency / Confusion Matrix:** A table that shows how well a model acts on test data for which the target values are known. As illustrated in Table 1, the diagonal indicates successfully identified occurrences and the off-diagonal shows classification errors.

| Class | 1 | 2 | ... | K |
|-------|---|---|-----|---|
| 1     | m_{11} | m_{12} | ... | m_{1K} |
| 2     | m_{21} | m_{22} | ... | m_{2K} |
| ...   | ...   | ...   |     | ...  |
| K     | m_{K1} | m_{K2} | ... | m_{KK} |

**Accuracy** \((TN + TP) / (TN + TP + FN + FP)\) computes a classifier’s ability to identify only the truthful cases for each class.

**Recall** \((TP / (TN + FP))\) is a classifier’s ability to find all truthful cases in a class.

**F1-score** \(2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})\): This shows a perfect balance because precision and recall are inversely related. When both recall and precision are important, then a high F1-score is useful.

A macro-average finds the metric independently for each class in order to calculate the average, whereas a micro-average computes the metric from the aggregate contributions of all classes. A micro-average is used for imbalanced datasets because it considers the frequency of each class. Micro-averaging and macro-averaging are methods for calculating a single number for each of the precision, recall, and other metrics across multiple classes.

### 7. Results and discussion

The proposed work investigates and compares the applicability of six different types of classifiers and nine significant sampling techniques (1. Random Oversampling, 2. Random Under sampling, 3. SMOTE, 4. ADASYN, 5. SVM SMOTE, 6. Borderline SMOTE, 7. SMOTENC, 8. SMOTEEN, and 9. SMOTETOMEK) to predict over four publicly available clinical datasets. Resampled datasets and original datasets were fed to six different classifiers one by one and assessed in terms of accuracy, precision, recall, and F1-score. During the model execution, the total size of the dataset was split into 60% instances as training data and 40% instances as test data. Moreover, k-fold cross validation is used to improve performance. Where k is the number of iterations, a model will be trained...
and validated wherein a different fold of data is held out for validation for each iteration and the leftover k-1 fold is often used for learning.

Each classifier has been tested under the following parameter settings:

**SVM:** Kernel: ‘linear’, max_iter: −1 (for no limit), gamma: ‘scale’ (kernel coefficient), decision_function_shape: ‘ovr’ (one-vs-rest), Degree: 3 (Degree of the polynomial kernel function (‘poly’)).

**GBN:** (Prior probabilities of the classes) priors: None, var_smoothing:1e-9.

**MLP:** random_state:1, hidden_layer_sizes:(10, 10, 10), max_iter: 10², solver: adam, activation: relu, learning_rate_init:10⁻³, learning_rate: constant, alpha: 10⁻⁴ (Regularization).

**k-NN:** n_neighbors:3, weights: uniform, leaf_size:30, p:2, metric: ‘minkowski’, algorithm: ‘auto’.

**Decision Tree (DT)** random_state:None, ccp_alpha:0.0, splitter: ‘best’, criterion: ‘gini’, max_depth:None, min_samples_leaf:1, min_samples_split:2, min_impurity_decrease:0.0, min_impurity_split: None, class_weight:None, max_features:None, min_weight_fraction_leaf:0.0, max_leaf_nodes:None.

Tables 2-5 contain the sample distribution of each class label before and after the application of class balancing techniques in the thyroid disease dataset, HCV disease dataset, breast tissue disease dataset, and skin disease dataset.

7.1. Performance comparison of different classifiers with different smote variants over thyroid disease dataset

In this section, we have expressed the results of the experiments performed over thyroid disease dataset. From Figs. 11–14, shows the six classifier’s performance measures in terms of accuracy, precision, recall and F1-score value with and without applying class balancing techniques over the thyroid dataset.

The Decision Tree (DT) with SMOTEEN was found superior in accuracy results, where it could classify above 96.9% of all the instances correctly, and so this combination has the highest measures in the thyroid dataset. In some cases, the accuracy may be the misleading claim, therefore, a better way to assess the performance of the classifier is to confirm it by precision (97.3%), recall (96.9%) and F1-Score (96.9%) also. Alternatively, sampling techniques ROS and ADSYN with DT also show comparable results (accuracy-94.3%, precision-95.2%, recall-94.3%, and F1-score of 94.3% approximately same) in the prediction of thyroid disease.

The classifier’s performance with the SMOTEEN class balancing technique in descending order came as: DT>svm>gnb>k-NN>lda>ann. Experimental result projected Decision Tree (DT) with SMOTEEN could distinguish (Hyperthyroid, Hypothyroid and Normal) 97.3% correctly and stood as the best performer whereas ANN could classify only 33.3% correctly and fell as the poorest performer in the case of thyroid disease prediction. Moreover, there is a significant 9% performance improvement in prediction when the prediction was done by DT with SMOTEEN and without class sampling technique.

7.2. Performance comparison of different classifiers with different smote variants over Hepatitis C Virus (HCV) dataset

This experimental demonstration was done to classify all four class labels (Portal Fibrosis—1; Few Septa—2; Many septa—3, Many septa—4) in Hepatitis C Virus (HCV) disease to identify the correct type of HCV disease. Based on the experimental results as presented in Figs. 15–18, the following conclusion was drawn- None of the classifiers could achieve acceptable classification performance in terms of precision, accuracy, recall and F1-score. The highest accuracy 41.2% is achieved by k-NN with SMOTEEN class balancing technique. This means its misclassification is 59.8% which recommends not to take the help of any of these models (DT, SVM, GNB, k-NN, LDA, and ANN) for prediction of HCV disease in patients. The possible reasons are hidden in the format of the dataset which contains 1385 instances with 29 features and four target classes. It has a diverse and large number of attribute values. The attribute value ranges have wide differences in minimum and maximum values (Features like RNA Base, RNA 4, RNA 12, RNA EOT, and RNA EF). In addition to this, range value fluctuation amongst different attributes is also affective learning by different models. This has caused overfitting in the model while learning over the HCV dataset. For the reasons stated earlier all the models underperformed. In other words, models are not able to generalize on a new dataset. Moreover, approximately a 17.6% increase in the performance of the k-NN model with and without using SMOTEEN class balancing techniques was observed.

**Table 2** Thyroid disease dataset: Samples distribution of each Label in Class.

| S.No. | Balancing Method         | Samples distribution of each Label in Class [Hyperthyroid→1; Hypothyroid→2; Normal→3] |
|-------|-------------------------|----------------------------------------------------------------------------------------|
| 1     | None                    | 1: 166, 2: 368, 3: 6666                                                               |
| 2     | Under Sampling          | 1: 166, 2: 166, 3: 166                                                               |
| 3     | ROS                     | 1: 6666, 2: 6666, 3: 6666                                                             |
| 4     | SMOTE                   | 1: 6666, 2: 6666, 3: 6666                                                             |
| 5     | ADASYN                  | 1: 6644, 2: 6704, 3: 6666                                                             |
| 6     | SVM SMOTE               | 1: 6666, 2: 6666, 3: 6666                                                             |
| 7     | Borderline SMOTE        | 1: 6650, 2: 6200, 3: 5314                                                             |
| 8     | SMOTENC                 | 1: 6666, 2: 6666, 3: 6666                                                             |
| 9     | SMOTEEN                 | 1: 6665, 2: 6616, 3: 6615                                                             |
| 10    | SMOTETOMEK              | 1: 6665, 2: 6616, 3: 6615                                                             |
Table 3
HCV disease dataset (Samples distribution of each Label in Class).

| S.No. | Balancing Method | Samples distribution of each Label in Class [Portal Fibrosis→1; Few Septa→2; Many septa→3, Many septa→4] |
|-------|------------------|--------------------------------------------------------------------------------------------------|
| 1     | None             | 1: 336, 2: 332, 3: 355, 4: 362                                                                   |
| 2     | Under Sampling   | 1: 332, 2: 332, 3: 332, 4: 332                                                                |
| 3     | ROS              | 1: 362, 2: 362, 3: 362, 4: 362                                                                |
| 4     | SMOTE            | 1: 362, 2: 362, 3: 362, 4: 362                                                                |
| 5     | ADASYN           | 1: 362, 2: 362, 3: 362, 4: 362                                                                |
| 6     | SVM SMOTE        | 1: 362, 2: 362, 3: 362, 4: 362                                                                |
| 7     | Borderline SMOTE | 1: 362, 2: 362, 3: 362, 4: 362                                                                |
| 8     | SMOTENC          | 1: 362, 2: 362, 3: 362, 4: 362                                                                |
| 9     | SMOTEEN          | 1: 12, 2: 13, 3: 10, 4: 7                                                                     |
| 10    | SMOTETOMEK       | 1: 233, 2: 238, 3: 224, 4: 223                                                                |

Table 4
Breast tissue disease dataset (Samples distribution of each Label in Class).

| S.No. | Balancing Method | Samples distribution of each Label in Class [car→1; fad→2; mas→3; gla→4; con→5; adi→6] |
|-------|------------------|------------------------------------------------------------------------------------------------|
| 1     | None             | 1: 42, 2: 30, 3: 36, 4: 32, 5: 28, 6: 44                                                    |
| 2     | Under Sampling   | 1: 28, 2: 28, 3: 28, 4: 28, 5: 28, 6: 28                                                   |
| 3     | ROS              | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |
| 4     | SMOTE            | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |
| 5     | ADASYN           | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |
| 6     | SVM SMOTE        | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |
| 7     | Borderline SMOTE | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |
| 8     | SMOTENC          | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |
| 9     | SMOTEEN          | 1: 34, 2: 20, 3: 14, 4: 27, 6: 34, 5: 30                                                   |
| 10    | SMOTETOMEK       | 1: 44, 2: 44, 3: 44, 4: 44, 5: 44, 6: 44                                                   |

Table 5
Skin disease dataset (Samples distribution of each Label in Class).

| S. No. | Balancing Method | Samples distribution of each Label in Class [Psoriasis→1; Seboreic dermatitis→2; Lichen planus→3; Pityriasis rosea→4, Cronic dermatitis→5, Pityriasis rubra pilaris→6] |
|--------|------------------|-----------------------------------------------------------------------------------------------|
| 1      | None             | 1: 112, 2: 61, 3: 72, 4: 49, 5: 52, 6: 20                                                     |
| 2      | Under Sampling   | 1: 20, 2: 20, 3: 20, 4: 20, 5: 20, 6: 20                                                     |
| 3      | ROS              | 1: 112, 2: 112, 3: 112, 4: 112, 5: 112, 6: 112                                               |
| 4      | SMOTE            | 1: 112, 2: 112, 3: 112, 4: 112, 5: 112, 6: 112                                               |
| 5      | ADASYN           | 1:112, 2:104, 3:112, 4:120, 5:113, 6:113                                                    |
| 6      | SVM SMOTE        | 1: 112, 2: 112, 3: 91, 4: 112, 5: 112, 6: 69                                                 |
| 7      | Borderline SMOTE | 1: 112, 2: 112, 3: 72, 4: 112, 5: 112, 6: 20                                                 |
| 8      | SMOTENC          | 1: 112, 2: 112, 3: 112, 4: 112, 5: 112, 6: 112                                               |
| 9      | SMOTEEN          | 1: 96, 2: 79, 3: 112, 4: 89, 5: 101, 6: 111                                                 |
| 10     | SMOTETOMEK       | 1: 112, 2: 112, 3: 112, 4: 112, 5: 112, 6: 112                                               |

Fig. 11. Accuracy of six classifiers with different balancing techniques for thyroid disease dataset.
Fig. 12. Precision of six classifiers with different balancing techniques for thyroid disease dataset.

Fig. 13. Recall of six classifiers with different balancing techniques for thyroid disease dataset.

Fig. 14. F1-score of six classifiers with different balancing techniques for thyroid disease dataset.
7.3. Performance comparison of different classifiers with different smote variants over breast tissue dataset

In the task of distinguishing six class labels (car→1; fad→2; mas→3; gla→4; con→5; adi→6) in breast tissue disease. The experimental evaluation has been conducted in terms of accuracy, precision, recall, and F1-score results are presented in Figs. 19–22, respectively. In combination with the SMOTEEN sampling technique, the highest accuracy achieved by different classifiers-DT, SVM, GNB, k-NN, LDA, and ANN is 99.9%, 92.2%, 90%, 89.1%, 82.8%, and 33.3%, respectively. Alternatively, ROS with different classifiers-DT, SVM, LDA, k-NN, GNB, and ANN obtained the second highest accuracies 94.3%, 84%, 72.6%, 71.7%, 61.3%, and 40.6%, respectively. Evidently, these classification models have also shown similar trends in the cases of precision (Fig. 16), recall (Fig. 17),
SMOTEEN improved the Decision Tree (DT) classifier’s accuracy by 10% when compared to the Decision Tree without any sampling technique.

When we compare the contributions of all nine class balancing techniques to improving accuracy, we discovered the following ordering from best to worst: SMOTEEN > ROS > ADASYN > BorderlineSMOTE > SMOTE > SMOTENC > Undersampling > SVM-SMOTE > SMOTETOMEK.

In this study, the Decision Tree (DT) classifier proved to be a promising classification model, with 96.9% ability to correctly classify and 97.3% trustworthiness in predicting multiple disease types in breast disease. In contrast to Decision Tree (DT) classifier, the ANN proved as the poorest classification model with 42.5% ability to correctly classify and 43.5% trustable in making multiple disease types.

and F1-score (Fig. 18) measurements.

Fig. 18. F1-score of six classifiers with different balancing techniques for HCV disease dataset.

Fig. 19. Accuracy of six classifiers with different balancing techniques for breast tissue dataset.

Fig. 20. Precision of six classifiers with different balancing techniques for breast tissue dataset.
Fig. 21. Recall of six classifiers with different balancing techniques for breast tissue dataset.

Fig. 22. F1-score of six classifiers with different balancing techniques for breast tissue dataset.

Fig. 23. Accuracy of six classifiers with different balancing techniques for skin disease dataset.
in breast disease prediction.

7.4. Performance comparison of different classifiers with different smote variants skin disease dataset

In the process of identifying six class labels (Psoriasis→1; Seboreic dermatitis→2, Lichen planus→3, Pityriasis rosea→4, Cronic dermatitis→5, Pityriasis rubra pilaris→6) in the skin disease dataset, the various classification models with different class balancing techniques have been trained and tested for their performance. Experimental evaluation of different models is demonstrated in terms of accuracy, precision, recall, and F1-score in Figs. 23–26, respectively.

In combination with the SMOTE sampling technique, the highest accuracy achieved by different classifiers-SVM, LDA, DT, ANN, k-NN, GNB, is 99.3%, 97.4%, 97%, 95.5%, 92.2%, and 88.5%, respectively. SMOTENC with LDA, SVM, and DT attained promising accuracies of 99.3%, 98.1, and 98.1%, respectively. In addition to this, SMOTENC with LDA, SVM, and DT also shows correspondingly better values of precision (Fig. 24), recall (Fig. 25), and F1-score (Fig. 26).

In the case of the skin disease dataset, all six classification models proved promising with different sampling techniques, and the overall performance of the model reached up to 99.3% accuracy, which is positively supported by precision, recall, and F1-score values. The reason for high performance lies in the multiclass dataset, which contains only 366 instances with 33 features of categorical nature. In general, due to the application of sampling technique, a 2% improvement in accuracy was observed in distinguishing the six class labels in the skin disease dataset. Except for the undersampling technique, all the other eight class balancing techniques gave similar influences on the performance improvement of all six classifiers under the study.

ROS (Random Over Sampling) improves binary classification by generating synthetic balanced cases from continuous and categorical data using conditional density estimates of the two classes. SMOTE is a data augmentation approach that generates synthetic data points by slightly moving a data item in the direction of its sibling. It ensures that the synthetic data point is not a perfect replica but also not too distinct from known minority class observations. For different minority classes that are more difficult to learn, ADASYN uses a weighted distribution. As a result, the categorization decision boundary is improved. The reason for adopting SVM-SMOTE to improve classification is because it is more concerned with data separation. Synthetic data is generated at random along the support vectors. Because Borderline-SMOTE only generates synthetic data along the decision boundary of two classes, it is more suited to binary classification. For datasets with mixed (categorical and continuous) features, SMOTE-NC (Nominal and Continuous) is more appropriate. SMOTE-ENN combines SMOTE’s ability to generate synthetic instances for minority groups with ENN’s ability to delete some extraneous features, resulting in improved classification accuracy. SMOTE Tomek is a hybrid approach that combines Tomek’s under-sampling with Tomek’s oversampling (SMOTE). Oversampling adds data to the minority class, whereas undersampling removes data from the majority class that is unnecessary or unimportant. As a result, a well-balanced dataset is formed, which increases machine learning algorithms’ classification accuracy.

8. Conclusion and future work

People with pre-existing health issues such as like thyroid disease, Hepatitis C Virus (HCV), breast tissue disease, Chronic Dermatitis etc., may be at a higher risk of developing deadly COVID-19 and its mutant variants. So, early, and correct identification of these diseases is essential. Machine learning algorithms can assist in the early and quick diagnosis of numerous diseases in a huge population like India. However, machine learning algorithms perform poorly when a dataset does contain a class imbalance issue. Smote versions are used to overcome the class imbalanced problem. The experimental work revealed that combining class balancing strategies with classification models enhanced the overall performance of all six classifiers over clinical datasets. The combination of SMOTEEN and Decision Tree often gave better results in distinguishing the class labels in clinical datasets (Thyroid, Hepatitis C Virus, breast tissue disease and skin disease). Consequently, this study can be taken as a key insight for guidance to identify patient and would
help the doctor in clinical decision support system. It advocates the recommended classification and class balancing technique with respect to specific data to able to detect the disease automatically with high accuracy. In our next study, we will investigate and build the most effective multi-class balancing method to handle the multi-class imbalance problem in data.

**Declaration of Competing Interest**

None.

**Data availability**

Data will be made available on request.

**References**

[1] WHO. Hepatitis C. World Health Organization; 2021. https://www.who.int/news-room/fact-sheets/detail/hepatitis-c.

[2] Dubey AK, Mohbey KK. Enabling CT-scans for COVID detection using transfer learning-based neural networks. J Bimol Struct Dyn 2022. https://doi.org/10.1080/07391102.2022.2034668.

[3] Schmid S. The good, the bad and the ugly – identification and characterisation of thyroid hormone-dependent biomarkers. University of Lübeck; 2020.

[4] Kratzsch J, Pulzer F. Thyroid gland development and defects. Best Pract Res Clin Endocrinol Metab 2008;22(1):57–75.

[5] Seeihan MT. Biochemical testing of the thyroid. TSH is the best and, oftentimes, only test needed—a review for primary care. Clin Med Res 2016;14(2):83–92.

[6] WHO. Breast cancer now most common form of cancer: Who taking action. World Health Organization; 2021. https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action.

[7] Basra MKA, Shahruih M. Burden of skin diseases. Expert Rev Pharmacoecon Outcomes Res 2009;9(3):271–83.
Vinod Kumar is Ph.D. (Computer Applications) from Maulana Azad National Institute of Technology- Bhopal (India) in 2019. He is working as Associate Professor in Dept. of CSE, KL Deemed to University, AP, India. His research interest is in the field of Big Data Analytics, Web Mining, Machine Learning and Blockchain Technology, and Internet of Things.

Gotam Singh Lalotra (Ph.D.) is an Assistant Professor in the Dept. of Computer Science at Govt. Degree College Basohli affiliated to the University of Jammu. He has the teaching and research experience of 8 years. He has published various research papers in reputed journals. He is interested in Computer Security, Machine learning and Data Mining.

Ravi Kant Kumar is an Assistant Professor in the Dept. of Computer Science and Engineering at SRM University, Andhra Pradesh, India. He received his Ph.D. in Computer Science and Engineering from National Institute of Technology Durgapur, India, in the year 2019. His areas of interests are: Image Processing, Computer Vision, Visual Saliency, Pattern Recognition, Machine Learning etc.