Artificial Intelligence & Robotics Development Journal
Volume 3, Issue 3, pp 248-268, July 2023, https://doi.org/10.52098/airdj.2023332
ISSN: 2788-9696 Received: 10/8/2023 Revised: 22/8/2023 Accepted: 23/8/2023

Machine Learning Approach for Detecting and Classifying the Cancer type using Imbalanced Data Downsampling

Farshad Kiyoumarsi[1, *] and Sara Wisam[2]
[1]Department of Engineering, Faculty of Computer, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran.
[2]Department of Engineering, Faculty of Computer, Isfahan Branch, Islamic Azad University, Isfahan, Iran.

* Corresponding author: Farshad Kiyoumarsi1,*, farshad.kiyoumarsi@iau.ac.ir

Abstract

One of the most important applications of medical data mining is the early diagnosis of diseases with high accuracy. In the meantime, timely diagnosis of cancer as one of the main causes of death is of special importance. However, the classification and diagnosis of cancer is challenging due to the unbalanced nature of related data. In the data related to cancer disease, there is usually a minority class (patient samples) and a majority class (healthy people samples), which diagnoses the disease from the minority samples, and this is a challenge for the classifiers. This work investigated the problem of classifying the imbalanced data related to cancer disease using a machine learning approach based on the K-Nearest Neighbor (KNN) clustering technique. In this method, the insignificant samples of the majority class are removed, and the data are balanced. The proposed method is simulated and evaluated on 15 cancer datasets selected from the general SEER database. The simulation results approve a high classification of cancer type based on the average detecting accuracy criterion of more than 90%. Moreover, the current result is more efficient and improves classification accuracy compared to the methods proposed by other researchers in the literature survey.

***Keywords**:* unbalanced data; Machine learning, cancer diagnosis; downsampling; K-means algorithm.

1. Introduction

The leading cause of mortality in the upcoming years will be cancer, a serious health concern that has recently seen a rise in incidence (Siegel et al., 2023). In 2020, cancer was considered one of the most significant causes of death in the international community, with a rate of nearly 10 million deaths (WHO, 2020). As a result, healthcare and therapy data analysis is essential to clinicians to predict diseases, diagnose them in their early stages, and make the best therapeutic judgments. The advantages of using clinical information for pathological cancer diagnosis include lowering medical waste, improving the effectiveness of medicine treatment, promoting early diagnosis, and cancer primary prevention. The likelihood of a successful therapy increases with early cancer diagnosis, before growth and metastasis (Sarkar, P., & Srivastava, 2022). However, it becomes nearly incurable in the advanced stages, and the likelihood of surviving cancer sharply declines. The distribution of class imbalances is a general issue for real-world medical data, especially cancer data. The unequal quality of the major and minority classes is one of the key issues with employing data gathering for diagnosis and therapy. A fundamental difficulty for learning classifiers is examining data sets with an unbalanced distribution of classes (Thölke et al., 2023). When there is an imbalance in the distribution of representative samples among the comparing classes, one class has fewer samples than the others. Typically, the primary class is viewed as the negative class, while the minority class is considered positive. The class mismatch is seen as a fundamental difficulty in machine learning. Problems caused in the real world by unbalanced data classification are of interest to researchers. This work tries to identify cancer from unbalanced data using downsampling based on KNN classification.

One of the most crucial and frequently discussed issues in classification problems is learning classifier from unbalanced data (Alkharabsheh et al, 2022). The Data mining approach successfully investigates many problems with enormous data clusters to determine association relations between variables and to predict future values perfectly (KHAMIS & Yousif, 2022). The data is saved in Corpus-based Machine Learning clusters that access vast health information resources, which is identified as a critical figure in all areas of Natural language processing applications (Yousif J., 2019). Because having to learn machines will gravitate toward the category with more samples by having trained on unbalanced data, the accuracy and effectiveness of the system will decrease when the sample number related to one classification is very small compared to the other category. In many databases, the main goal is to detect specimens that are few but also have high importance and are unbalanced. Unbalanced data exist in various fields, including network intrusion detection, fake internet user detection, disease diagnosis, information retrieval, and text

classification (Siegel et al., 2021). Training data is required for the system to be highly precise and reliable. When existing classes are distributed equally, a class crucial to the application domain frequently has fewer instances than the class that makes up the majority. Unbalanced describes this collection of data (Yang & Chen, 2015). The performance of the proposed classification system will be lower when the learning is based on data sets.

This work proposes a novel approach to classify unbalancing data based on the KNN technique. Also, to efficiently provide a machine learning-based approach for analyzing unbalanced medical data and cancer diagnosis.

2. Literature Survey

The three-classification approach and the competition double learning model (TSFS-TCEM), a basis of article for selective feature extraction and group learning, has been presented for the diagnose in the study (Tang, et al., 2021). The project's authors develop multicomic information on breast cancer in the first phase by combining functional transcript and spectrometry data. This is the first time they have employed this combination of biological data to detect breast cancer. The second step of the suggested strategy involves the introduction of different models during selecting features and the building of the diagnostic model. The issue of data imbalance brought on by a single classifier is resolved by the three-step feature selection process, which incorporates features from several data kinds. The diagnostic accuracy of TSFS-TCEM is 99.64 percent, which is higher than all other approaches. Additionally, the sensitivity, specificity, and F-Measure technique are above 99.63 percent in the 5-fold cross-validity study. To get beyond the aforementioned issues and boost classification accuracy, study (Gan et al., 2020) proposes an integrated TANBN with cost-sensitive classification method (AdaC-TANBN). The AdaC-TANBN algorithm executes the classification for imbalanced data in medical diagnosis by first applying the variable error classification cost, which is defined by the probability of sample distribution, to the train classifier. The Cleveland Heart Database (Cardiac), Indian Liver Patient Database (ILPD), Dermatological Database, and Cervical Cancer Risk Factor Database (CCRF) from the UCI Learning Repository have all been examined to determine the efficacy of the proposed approach. According to experimental findings, the AdaC TANBN algorithm can outperform other cutting-edge comparison techniques. According to study (Liu et al., 2020), a new group learned model with three stages—data preprocessing, fundamental training classifications, and final set—is proposed for medical diagnosis using unbalanced data. By combining the Artificial Minority Oversampling (SMOTE) technique with the CVCF filtering technique, which can not only create minority samples and thereby input samples, we introduce the extension of the SMOTE technique in the initial stage of data preparation. for balance in order to perform well during the classification process, filter out

noisy data as well. It uses the Group Support Vector Machine (ESVM) classification method during the classification stage, which has the advantages of great generalization performance and classification accuracy because it is built from many SVM classification structures. Additionally, the balanced majority vote strategy and the simulated annealed genetic algorithm (SAGA) are introduced in the final phase of the group strategy in order to maximize the weight matrix and hence enhance classification performance. On nine unbalanced medical datasets, the effectiveness of the suggested group learning approach was examined. The experimental results unmistakably proved that the proposed learning model is better to previous advanced classes. An imbalanced sampling strategy using self-learning (ISPL) is suggested in the work (Wang et al., 2020) to efficiently choose top quality samples to increase resilience. According to experimental findings, the proposed ISPL method boosted classification accuracy by about 16 percent in comparison to the typical yield produced using other sampling techniques. Additionally, the novel technique was successful in choosing a few significant genes for additional study.

The Learning Paradigm Using Unique Information (LUPI) and or the Maximum Mean Difference (MMD) frameworks are integrated inside a novel Double-Surveillance TL Network (DDSTN) proposed in Paper (Han et al., 2020). The suggested approach can undertake further supervised transfer among unpaired data in addition to fully utilizing shared tags to efficiently direct knowledge transmission through the LUPI paradigm. The majority of the MMD requirements for enhancing knowledge transfer are introduced in this work. According to experimental findings on a dataset of breast ultrasound images, the suggested DDSTN outperforms all other sophisticated equivalent BUS-based CAD methods. The study (Xie et al., 2020) looked at the impact of re-sampling methods for unbalanced data sets inside the PET-based radiographic prediction model in patients with head and neck cancer (HNC). Two patient groups underwent radiographic analysis: 182 patients with HNC from an extensive database and 166 patients at the Research Center who had just received a nasopharyngeal carcinoma (NPC) diagnosis. For correlation study of overall survival (OS), disease-free survival, typical PET values and potent radiography features were retrieved (DFS). In order to predict survival, researchers have examined a mixture of 10 resampling techniques (over-sampling, under-sampling, and combination-sampling). Test kits for maintenance were evaluated for their diagnostic performance. Utilizing Monte Carlo cross-validation and Nemenyi post hoc analysis, statistical differences were examined. With minimal loss of F-measures in the classifier, over sampling techniques like ADASYN and SMOTE can enhance predictive results in terms of G-mean & F-measures with in minority class. For their NPC group, the researchers found the best radiographic PET operating system prediction model (AUC 0.82, average G 0.77). When it was checked on an external

data set, similar results showing that over-sampling additional effects predictive performance were seen, indicating generalizability. predicted the uneven data set and displayed it. In order to facilitate easy copying in future studies, they have also created an innovative source solution for automatic vehicle computations and comparisons of various resampling technics and machine learning classifiers.

The authors of (Alam et al., 2020) have mostly concentrated on examining the potential risks for MM. Data from pleural patients were used to determine the symptoms of the disease. The dataset, however, only includes mesothelioma patients in good health. The database is prone to imbalanced class issues where there are significantly fewer MM patients than healthy people. The artificial sampling approach of the synthetic minority has been utilized to solve the class imbalance issue. Exploring the link law for a which was before dataset forms the foundation of the Apriori algorithm. Before applying the Apriori algorithm, features that were redundant and useless were deleted. Additionally, communication rules were developed in the data set and numerical properties were divided into nominal properties. The findings of this study indicate that the key risk factors for MM are ESR rate, asbestosis and length, and the ratio of thoracic to serum lactic dehydrogenase. Early identification and management for the condition can help to avert the disease's more severe phases. The risk of various diseases, including as heart disease, psychiatric disorder, diabetes, and anemia, might rise if risk factors are not identified.

In order to address the issue of EISM data, Work (Fujiwara et al., 2020) suggests a brand-new technique known as amplification mixed with exploratory sampling under distribution-based sampling (HUSDOS-Boost). HUSDOS-Boost removes extra majority class based on earlier amplification findings and generates false minority samples based on a minority classification process in order to produce an artificially be so from the actual unbalanced dataset. Eight unbalanced datasets were used to test the functionality and performance of HUSDOS-Boost. In order to detect patients with stomach cancer, this algorithm was also used to significant clinical human resource data. These findings demonstrate that HUSDOS-Boost outperforms existing unbalanced data management techniques, particularly when using EISM data. As a result, the suggested HUSDOS-Boost is an effective way for assessing data related to human resources.

For the first time, a thorough investigation of the effects of the issue of data sets on the information of cancer sufferers has been carried out in the references article (Siegel et al., 2021). In this case, the two primary balancing strategies employing 18 algorithms are over-sampling and under-sampling. While the followings are CNN, CNNTL, NCL, OSS, RUS, SBC and TL, over-sampling techniques include ADASYN, ADOMS, ROS, Safe-Level-SMOTE,

AHC, Borderline-SMOTE, SMOTE, SMOTE-TL, SMOTE-ENN, SPIDER and SPIDER2. In order to investigate how balancers affect classifier performance, four classifiers—RIPPER, MLP, KNN, and C4.5—are utilized as learners. The 15 cancer datasets from the SEER program that were utilized for this investigation also include datasets for kidney, soft tissue, rectum, prostate, colon, bone, larynx, breast, bladder, cervix, larynx, melanoma, thyroid, testis, and lip. The results of this study concentrate on investigating how class imbalance affects classifier performance, comparing the overall effectiveness of preprocessing strategies, and classification all data, and finally identifying the best balance and classifier for every type of cancer data. tuning. The findings show that utilizing balancers leads to a noticeable improvement. After using balancing strategies, the accuracy of the different classifiers in the imbalanced cancer data has improved with the AUC measurement in 90% of the cases. Friedman tests are employed to increase accuracy, and it's fascinating to note how each form of cancer data responds differently to various balancing and categorizing techniques. The sampling balancing approaches also prove to be more effective than the selected samples when the mean rank of each approach and the classifier employed for the data set are taken into account. Table 1 presents a summary of used methods and datasets.

3. Background about Cancer

The term "cancer" refers to a sizable family of illnesses that include abnormal cell growth and the capacity of these cells to spread to and harm other areas of the body. Subsets of neoplasms develop during this disease; a neoplasm, also known as a tumor, is actually a collection of cells that have grown in an uncontrolled manner to form what is typically a mass, but this material can spread throughout the body. This illness can begin in any organ, including the blood, skin, digestive, urinary, bone, muscle, or eye systems, and in more late stages, this could spread throughout the body. The symptoms will vary depending on where in the skin the mass is located. Cancer is caused by a number of factors, including genetics, toxins, radiation, and others. There are numerous cancer treatment options, each of which will vary depending on the type and extensiveness of the disease's progression; these options either result in a cure or slow the disease's progression (Miller et al.,2022). There are numerous approaches, including surgery, chemo, radiation, drug therapy, etc.; When treating a particular type of cancer, a combination of currently available treatments may occasionally be required. Surgery is typically used in the early stages of cancer. Disease is a multi-stage disease where cells lose their capacity to divide and develop normally as a result of abnormal growth or death. This results in the invasion, mutilation, and corruption of normal tissue. Tumors are formed as a result of the collection of these cancer cells and the mutilation of healthy tissues.

Table 1: A summary of related research about the cancer

Researcher & Year	Method Used	Datasets Used	Findings
Tang et al., 2021	Three-classification approach and TSFS-TCEM	Breast cancer data	Diagnostic accuracy of TSFS-TCEM is 99.64% for breast cancer detection.
Siegel et al., 2021	Over-sampling and under-sampling with various balancing strategies and classifiers	Multiple cancer datasets (e.g., kidney, soft tissue, rectum, etc.)	Balancing strategies significantly improve classifier performance for imbalanced cancer data.
Gan et al., 2020	Integrated TANBN with AdaC-TANBN	Various medical datasets (Cleveland Heart, Indian Liver, Dermatological, Cervical Cancer Risk Factor)	AdaC-TANBN outperforms other techniques in medical diagnosis.
Liu et al., 2020	Group learning model with SMOTE and ESVM	Unbalanced medical datasets	Group learning approach performs better than previous methods on unbalanced medical datasets.
Wang et al., 2020	Imbalanced sampling strategy using ISPL	Not specified	ISPL method increases classification accuracy by about 16% compared to other techniques.
Han et al., 2020	Double-Surveillance TL Network (DDSTN)	Dataset of breast ultrasound images	DDSTN outperforms other methods for breast ultrasound image analysis.
Xie et al., 2020	Various resampling techniques for PET-based model	Radiographic data for head and neck cancer (HNC)	Over-sampling techniques like ADASYN and SMOTE improve predictive results, especially in the minority class.
Alam et al., 2020	Synthetic minority sampling, Apriori algorithm	Dataset of pleural patients	Key risk factors for mesothelioma include ESR rate, asbestosis, and others.
Fujiwara et al., 2020	HUSDOS-Boost for handling EISM data	Various unbalanced datasets (including clinical human resource data)	HUSDOS-Boost outperforms other unbalanced data management techniques, especially with EISM data.

The tumor is innocuous or non-cancerous if it only has a few layers and does not expand to other tissues or organs. Malignant or cancerous refers to a tumor that spreads or has the ability to spread and encircle other tissues and organs. Some cancers metastasize, which means they take on aggressive attributes and spread to other parts of the body, primarily through blood and lymph, and produce new tumors (Siegel et al., 2021). Some individuals may also possess powerful immune systems that can control or eliminate cancer cells, both present and potential. There is data to suggest diet has a significant impact on the immune system's ability to stop the spread of cancer cells in the body. Also, other

variables such as fatness, lack of physical activity, systemic inflammatory and hormone levels can play a significant role in cancer (Hesketh R., 2023).

There are many more than 200 different types of cancer, each one named for the affected organ or tissue. For instance, brain cancer begins with brain cells and lung cancer with lung cells. The following cancers are included in the general classification (Elwahsh et al., 2023; Hesketh R., 2023).

- Skin, respiratory, colon, pancreas, and ovarian cancer are all examples of carcinomas, a type of cancer that develops in the tissues that encompass internal organs.
- Sarcoma: A form of cancer that develops in tissues related to bone, cartilaginous, fat, muscle, capillaries, etc.
- Leukemia: This form of cancer develops in blood-forming tissues, such as the bone marrow, and leads to the production of a significant number of platelets in the blood, which then circulate in the body.
- Cancerous lymphocytes are the site of the beginning of lymphoma (T or B cells). These white blood cells are involved in disease defense and are a component of the immune system. abnormal lymphocytes in lymphoma cancer They are produced in the body's lymphatic system, lymph nodes, and other locations.
- Melanoma is a cancer that develops in precursor cells to melanocytes. Specialized cells called melanocytes produce melanin. Melanomas typically develop on the skin, and so they can also grow in other tissues, like the eye.

In addition to diagnostic tests, performing physical exams, physical tests, and reviewing the patient's medical history are all necessary for cancer diagnosis. Cancer can be stopped from spreading if detected early enough. There are numerous diagnostic tests available to confirm or rule out the presence of cancer, as is obvious from the fact that no single test is capable of accurately diagnosing the disease. There are several methods used to diagnose cancer (Hesketh R., 2023) including Lab examinations, Genetic tests, Adoscopy, diagnostic imaging, and Tumor samples.

The initial steps in making a cancer diagnosis include a medical examination, health information, and a review of past symptoms. Depending on the cancer type or the affected area, tests may occasionally be carried out. Imaging is another effective way to find bodily anomalies. X-rays, CT scans, MRIs, and sonograms are frequently used equipment for body examination (Lanjewar et al., 2023). Endoscopy can also be used to assess the health of the stomach, throat, intestines, and other organs. During a biopsy or sampling, a cancer's type, stage, and extent are all identified in addition to the disease's diagnosis.

4. Data Mining Processing

Data mining is the process of automatically extracting knowledge from the massive quantities of data. For various applications, the word "usefulness" has different connotations (Aljabri & Yousif, 2023). In business, useful information will help the manager must make strategic decisions. In the case of geomorphological science, useful

information gives information on changes in the Soil system for a long time. Information retrieval in dataset is the process to convert raw data into valuable knowledge (Dubey S, 2022). The Knowledge discovery process includes several phases that could contain several sub-phases, as shown in Figure 1.

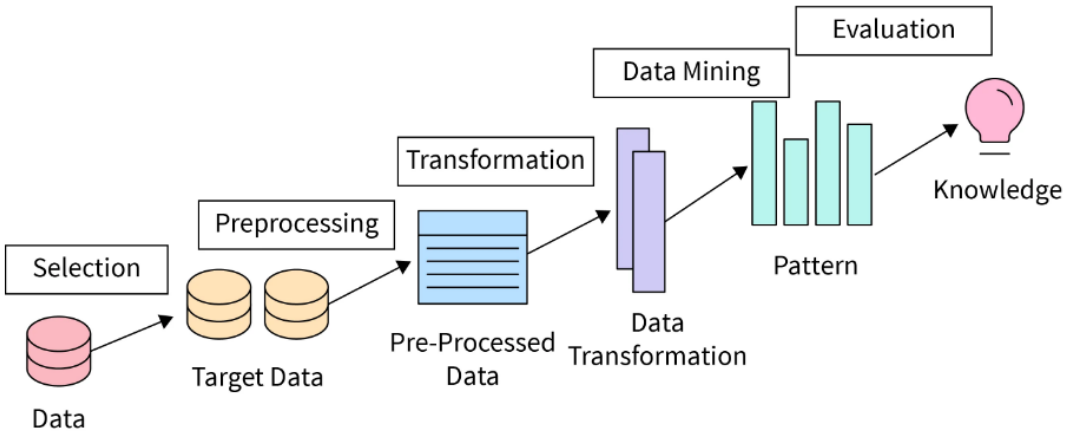

Figure 1: Knowledge discovery process in the database (Jodha, 2023)

In tasks requiring prediction, specific data features are employed to foretell the value of many other features. Descriptive tasks look for patterns in the data that can be explained by a human (Jodha, 2023). Classification and regression are two categories that can be applied to predictive tasks. Preprocessing entails operations to transform data into a format suitable for analysis. Since the format of data storage in databases is different, a lot of time is usually spent on data preprocessing Data preprocessing includes Data cleaning, Data integration, Data transformation, Data reduction (Sun et al., 2009; Yousif J, 2011). Data analysis is used to model and forecast the value of a variable, whereas classification is being used to forecast a discrete value. The problem of classification in pattern recognition is crucial. Decision trees, artificial neural, Bayesian networks, and closest neighbor are a few examples of categorization algorithms (He & Shen, 2007). The classification involves learning the target function f, which links the collection of characteristics x to the initial class label an. Sets of records make up the input data for classification. A feature of this example is specified by the binary (x, y), where x denotes the selection of characteristics and y denotes the class label. Data Post-processing is carried out to comprehend the outcomes of data mining.

4.1. Imbalanced Data Set

A data set is technically considered unbalanced if the distribution of its classes is uneven. Despite the fact that cases with extreme imbalances are what are typically meant by "unbalanced data." There are two types of imbalances

that can exist in a data set: imbalance between classes, in which some classes have much more samples than the others (Pearson et al., 2003; Weiss G., 2004). and a difference within a class where some subsets have more samples than others. compared to other subsets within the same class, has a lot less as shown in Figure 2 (Engati, 2023).

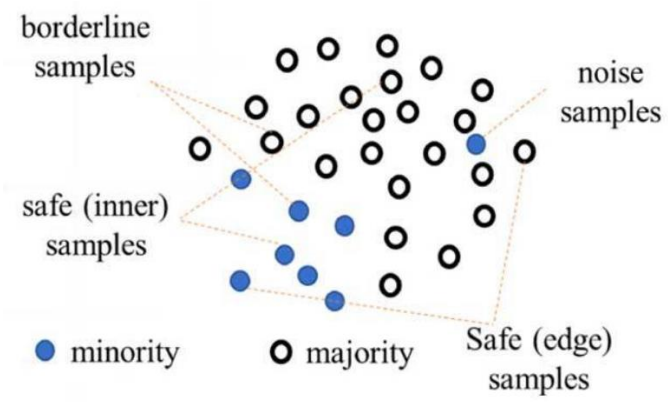

Figure 2: A: Data set with interclass imbalance (Engati, 2023).

4.2. Problem with imbalanced data classification

The biggest issue resulting from unbalanced data classification is the misdiagnosis of the minority class, which is more susceptible and likely than the majority class (Longadge & Dongre, 2013). Machine Learning algorithms are biased in favor of classes with more samples in this way (He & Garcia, 2009). The following issues with unbalanced medical data prevent classifier learning:

- Samples with little or no information and sparse training data: classification modeling and separating minority samples from the majority, more data and more formation are helpful (Krawczyk B, 2016).
- Class coincides: It is hard to apply discriminatory rules in this case, and instances of the minority class could be classified under so many general rules. Any basic classifier may learn the proper categorization if there is no category overlap.
- Brief breaks: It occurs when the minority class's concept is broken down into smaller concepts. Due to unbalanced instances of the classes, these sub-concepts make the situation worse.
- The effects of noisy data on imbalanced data: noise is a very tiny cluster of the minority class that may be ignored by the classifier (López et al., 2013). The learning process may be influenced by evaluative criteria that lead to sample ignorance (which is thought to be noise).
- The value of borderline samples in determining the allowable gap between positive and negative classes (López et al., 2013).

Borderline samples are found in the area surrounding the class boundaries where the majority and minority classes overlap. Samples near the borderlines are more likely to be misdiagnosed than samples farther away from them, making them more crucial for classification (kotun et al., 2022).

4.3. Background Of KNN Algorithm

The learning group uses the supervised K Nearest Neighbor algorithm for classification (the most frequent application) and regression. It is a flexible algorithm that can also be used to resample data sets and compute missing values. K nearest neighbors is taken into account, as the name suggests, in order to predict the class or constant value for a new data point (new data). The K nearest neighbors' classification algorithm is depicted in Figure 3 (Almomany et al., 2022; Tang et al., 2021).

Input: X: training data, Y: class labels of X, K: number of nearest neighbors.
Output: Class of a test sample x.

Start
Classify (X,Y,x)
1. ***for*** each sample x ***do***
   Calculate the distance: $d(x, X) = \sqrt{\sum_{i=1}^{n}(x_i - X_i)^2}$
   end for
2. Classify x in the majority class: $C(x_i) = argmax_k \sum_{X_j \in KNN} C(X_j, Y_k)$

End

Figure 3: Steps of K nearest neighbors' classification algorithm (Almomany et al., 2022)

5. The proposed method

One of the basic problems in medical data mining is unbalanced data analysis. This problem is very challenging in cancer data because it reduces the accuracy of classification and identification of patients and has a serious impact on early diagnosis and timely treatment of patients. In this case, the minority class is related to cancer samples and the majority class is related to healthy samples. In this work, a solution to the problem of cancer disease data imbalance and high accuracy classification of this disease has been discussed. The proposed method is based on the KNN algorithm using downsampling technique for classifying the imbalanced data. In an unbalance data set, the samples are not evenly distributed in the classes, so that the samples of one class are much more than the other class. The class that has more samples is called the majority class and the class that has fewer samples is called the minority class. In this work, we use a down sampling method using KNN method.

5.1. Downsampling of Data with KNN Algorithm

In pre-processing, the main action is to normalize the features extracted from the data. Normalizing features means placing them in a fixed range such as [0,1]. This problem makes the effect of all features to be considered in the same classification model. The following relationship is usually used for normalization as defined in equation 1.

$$x_N = \frac{x_i}{x_{max}} \quad \ldots (1)$$

After balancing the data, we train the model using the KNN algorithm. This algorithm classifies the data into several classes. The reason for using this algorithm is the ability to learn with high accuracy based on the use of nearest neighbor data, which increases the accuracy of classification and final diagnosis. Random sampling removes data samples from the majority class at random, so important information belonging to the majority class that is present in some samples may be missed. Therefore, in this research, we used the K-Means clustering method as a downsampling technique as shown in Figure 4.

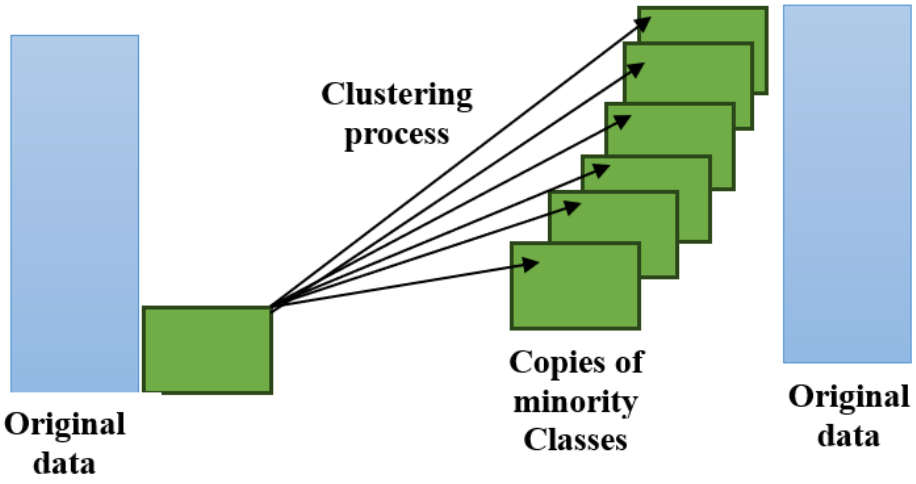

Figure 4: Downsampling technique

This alternative method for downsampling is proposed to preserve class samples that are more important and remove class samples that are less important in the data set in the model training phase. The purpose of the downsampling approach based on K-Means clustering is to maximize the accuracy of diagnosing cancer patients to deal with the problem of class imbalance. Our main goal in this work is to reduce the gap between the data samples belonging to the candidate class for cancer patients and the data samples belonging to the candidate class for healthy people by removing most of the data samples belonging to the candidate class of healthy people, which are less important in the dataset. The downsampling technique based on K-Means clustering is as follows:

- At first, we retrieve the majority class data samples in a separate data frame and determine the optimal number of clusters using cluster validity indices.
- Then the K-Means algorithm is performed on the majority class data frame. It should be noted that K is equal to the optimal number of clusters obtained in the previous step.
- Data samples that are similar to each other are grouped in clusters. A sample that has a high similarity between the center and itself is likely to be more important than a sample that is less similar to the cluster center. Our goal is to remove the data points far from the center because they are less important and keep the data points close to the center.

Hence, based on this concept, we calculate the Euclidean distances between the data points and their respective centroids in each cluster. To remove the data points that are far from the center, first we normalize the distances according to the total number of data points in each cluster, where the normalized data points (DN) in each cluster are calculated as in equation 2.

$$D_N = \frac{D_C}{N_C} \quad \dots (2)$$

In the above relation, D_C is the distance between the data point and the center of the cluster and N_C is the total number of data points in the cluster. Then we calculate the average of the data points in each cluster and name them $m_1, m_2, \dots, m_k$, where K is equal to the number of clusters or cluster centers, and the average for each cluster is calculated as in equation 3.

$$m_K = \frac{S_C}{N_C} \quad \dots (3)$$

In the above relationship, S_C is the sum of the distances of all data samples belonging to the cluster, and N_C is the total number of data points in the cluster. To remove data samples that are far from the center, we choose a threshold value. The threshold value is selected by calculating the mode from the average obtained in equation (2). The reason for choosing the mode from equation (2) is that it is suitable for our problem and we have obtained discrete and compact clusters using the mode over the mean. The threshold value is calculated as in equation 4.

$$Threshold = mode(m_1, m_2, m_3, \dots, m_k) \quad \dots (4)$$

Finally, after calculating the threshold value, the data samples whose distance from the center is greater than the threshold value are considered insignificant and hence are removed from the dataset. After balancing the data and removing some samples from the majority class, it is time to classify the data and identify which class each sample belongs to (sick or healthy). In this work, k-nearest neighbor algorithm is used for data classification. KNN algorithm is a strong classification algorithm based on Euclidean distance. Test data should be used to evaluate the proposed

method. First process is the data balancing and classifier training to be done using the training data. But to evaluate the proposed method, finally, the classification model is evaluated on the test data. In this work, we selected 70% of the database samples for training and balancing, and we considered the remaining 30% for testing the proposed algorithm.

6. Evaluation of the Results

This section presents the evaluation criteria for assessing the proposed method and describes the used datasets. In addition, simulation results are compared as evaluation criteria for the proposed method to examine the quality of classifying the training datasets. Similarly, a comparison of the performance of the proposed method with the other studies is implemented to prove the superiority of the current proposed method.

6.1. Database

In this work, we use the database Surveillance, Epidemiology, and End Results (SEER) as a reliable and unique resource to investigate different aspects of cancer (Shen et al., 2022) that combines patient data at the level of cancer region, patient pathology, disease stage and cause of death. This database is developed and maintained by the US National Cancer Institute (NCI), an authoritative data source for cancer incidence and patient survival in the US population. The SEER office collects patient data regularly. Data collected by SEER include demographics, primary cancer site and tumor morphology, stage at diagnosis, first course of treatment, and vital signs follow-up and patient survival information. Death data, reported by SEER, are provided by the National Center for Health Statistics. The dataset, collected from 1973 to 2014, includes 133 cancers. The data set related to 15 types of cancer used in this work is shown in Table 2 that includes the number of samples, the number of minority class samples, and the number of majority class samples.

6.2. Evaluation Criteria

Different evolution approaches were used to examine the accuracy and efficiency of simulated and predicted results (Kazem et al., 2022; Yousif et al., 2022). To compare the work with the reference article, the evaluation criteria used in the article, i.e., the AUC criterion, should be used.

The AUC criterion is defined as in equations (5).

$$AUC = \frac{1+TP_{rate}-FP_{rate}}{2} \quad \ldots (5)$$

where, TP_{rate} shows how many percent of minority class samples are correctly classified and FP_{rate} shows how many percent of majority class samples are wrongly classified. The higher is the AUC, the better the classification performance is.

Table 2: The 15 types of cancer used in this work

	Data set	No. of instances	No. of minority (non-survival)	No. of majority (survival)
1	Kidney	1053	459	594
2	Soft tissue	10,697	4191	6506
3	Bladder	12,324	4514	7810
4	Rectum	11,730	4113	7617
5	Colon	8171	2845	5326
6	Bone	3782	1233	2549
7	Larynx Glottic	1029	321	708
8	Breast	26,092	4974	21,118
9	Cervix	18,111	2850	15,261
10	Prostate	18,933	2243	16,690
11	Oropharynx	1211	140	1071
12	Melanoma	13,959	1596	12,363
13	Thyroid	13,817	867	12,950
14	Testis	4921	269	4652
15	Lip	1465	54	1411

Also, in this research, the criteria of accuracy, recall and F-score were also used, which are defined as in the equations (6, 7, 8, 9).

$$Accuracy(acc) = \frac{TP+TN}{TP+TN+FP+FN} \quad \dots (6)$$

$$precision = \frac{TP}{TP+FP} \quad \dots (7)$$

$$recall = \frac{TP}{TP+FP} \quad \dots (8)$$

$$F1score = 2 * \frac{recall*precision}{recall+precision} \quad \dots (9)$$

where TP means true positive predictions about patient survival, TN true negative predictions about patient death, FP false positive predictions about patient survival and FN false negative predictions. It is about the death of the patient.

6.3. Results of the proposed method

In this section, we will examine the simulation results of the proposed method in classifying data related to 15 types of cancer, all of which are unbalanced data types. The main goal is a two-class classification, the first class is related to the samples that survived the disease (majority class) and the second class is the deceased samples (minority

class). As explained in the proposed method chapter, in the simulation after data preprocessing, the data should be balanced by K-means clustering and then the classification is done by KNN algorithm. In Table (3), the name of each 15 different data, each of which is related to a type of cancer, is presented with the values of the evaluation criteria obtained by the proposed method (K-Means - KNN). For example, the classification accuracy for colon cancer is 93.02%.

Table 3: The results of the proposed method on the different data available in the SEER data in the form of evaluation criteria

Data Name	F- Score	Precision	Recall	AUC	Accuracy
Kidney	0.876893	0.783986	0.994905	0.62612	0.781637
Softtissue	0.954318	0.927048	0.987758	0.713524	0.917711
Bladder	0.975561	0.952381	1	0.97619	0.954955
Rectom	0.969919	0.941667	1	0.970833	0.945736
Colon	0.957609	0.918919	1	0.959459	0.930233
Bone	1	1	1	1	1
larynx glottis	0.999512	1	0.999025	0.993333	0.999089
Breast	1	1	1	1	1
Servix	0.973224	0.948653	0.999118	0.971801	0.955267
Prostate	0.945055	0.898148	1	0.949074	0.914729
Oropharynx	0.955937	0.916667	1	0.958333	0.930233
Melanoma	0.968787	0.947182	0.992593	0.918036	0.943524
Thyroid	0.942377	0.914984	0.971982	0.919613	0.91738
Testis	0.98495	0.97037	1	0.985185	0.973064
Lip	0.925481	0.933333	0.929825	0.943678	0.94697

Figure (5) presents bar chart diagrams of the accuracy measure for different data by the proposed method. Figure 6 shows a Bar chart diagram of the measure AUC for different data by the proposed method. The tested and evaluated Accuracy and AUC for all SEER database data by the proposed methods are better than other studies. Figure (7) also shows the average evaluation criteria for all 15 data related to the database. For example, the accuracy of the proposed method in the classification of all 15 cancer diseases is equal to 94.07% on average. Also, the other evaluation factors, precision, recall, and F-score of the proposed methods are achieved at a high rate.

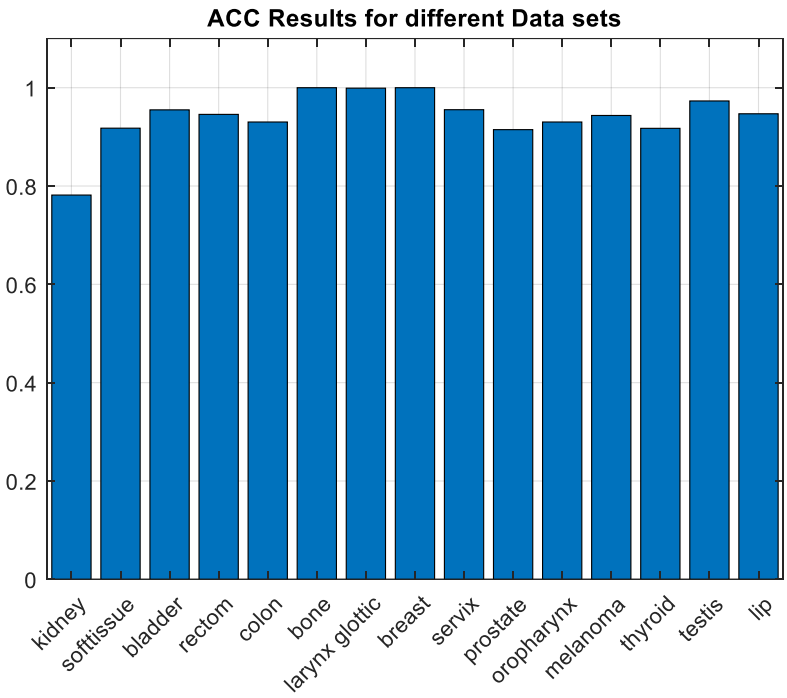

Figure 5: Bar chart diagram of the measure accuracy for different data by the proposed method

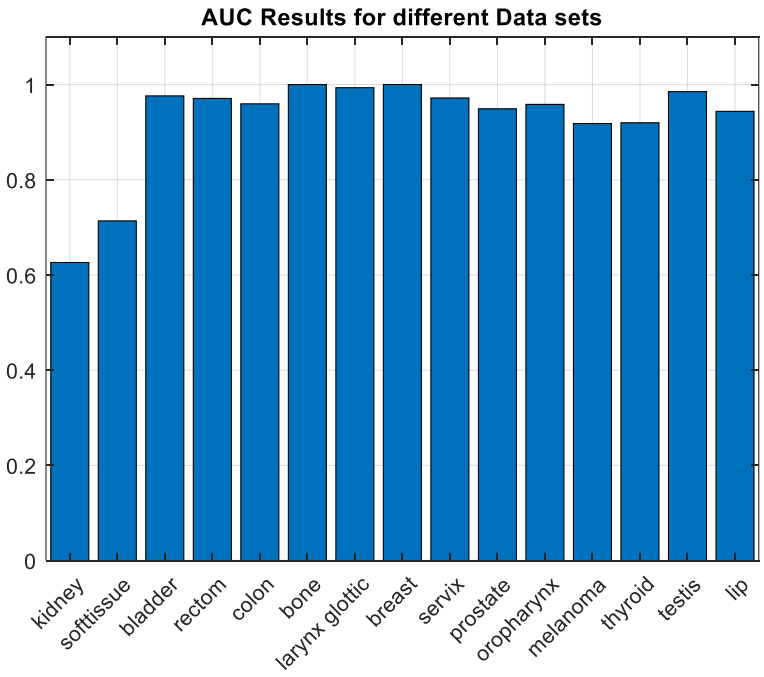

Figure 6: Bar chart diagram of the measure AUC for different data by the proposed method

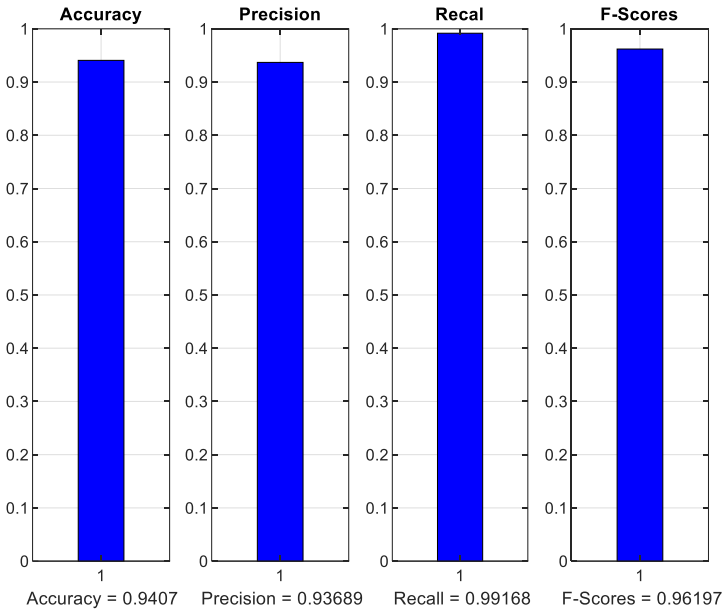

Figure 7: Numerical values of evaluation criteria averaged over 15 different data by the proposed method

Finally, in this section, the performance of the proposed method in the classification of 15 data related to different cancers, all of which were unbalanced data, has been compared with the performance of the methods in the reference article (Fotouhi et al., 2019). This comparison is done in the form of AUC criterion, and it should be noted that the values mentioned for the proposed method, in addition to being the average of 15 different data, are also averaged from 20 different simulation runs to increase the validity of the results. That is, in each simulation run, the average AUC is taken for 15 data, and finally, the average value of 20 different runs is presented in Table 3 and presented in Figure 8.

Table 3: The average AUC criterion for the proposed compared to other studies

Method	**Average AUC measured**
C4.5	75.4
KNN	68.7
MLP	66.9
RIPPER	77.5
The proposed method (K-means-KNN)	**92.44**

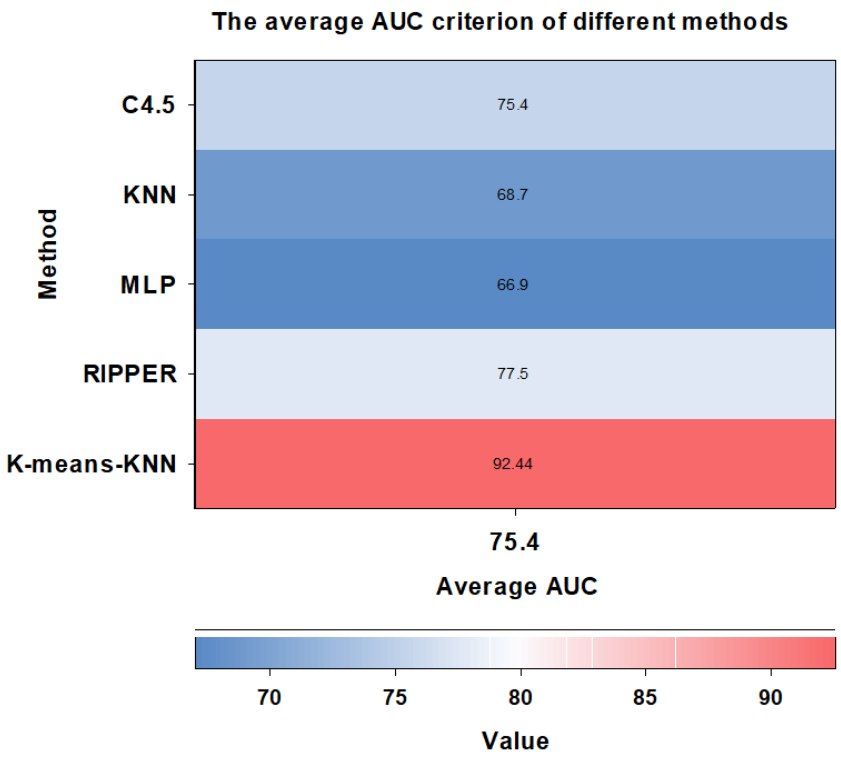

Figure 8: The average AUC of different used methods

7. Conclusion and Suggestions

Due to the increase in cancer patients, this disease will be the first cause of death in the world in the next few decades. As a result, the analysis of medical data related to this disease and the prediction and early diagnosis of this disease significantly impact patients' treatment process and survival. The benefits of clinical information for diagnosing cancer in the pathological stage include saving medical resources, increasing the efficiency of drug treatment, encouraging early diagnosis and primary prevention, and reducing medical waste. Imbalance in samples of different classes in medical data is a general problem for real-world medical data, especially cancer data. Analyzing datasets with unbalanced class distribution is a fundamental challenge for learning classifiers. The misdiagnosis of the minority class, which is more likely and sensitive than the majority class, is the biggest problem caused by unbalanced data classification. In most cases, algorithms assume a balanced distribution of classes in classification problems. As such, these algorithms are inefficient in handling the complex imbalanced datasets prevalent in the real world, especially in the medical field.

In this work, a scheme based on K-means clustering is presented to balance the distribution of classes related to 15 different types of cancer data, which were all unbalanced, to increase the classification accuracy of these data. In

the proposed method of this project, the insignificant samples of the majority class are first removed by the K-means clustering method. Then the balanced data are classified by the KNN classification algorithm. Based on the simulation results of the proposed method, the average criterion AUC for 15 different data is equal to 92.44, which has increased compared to other methods in the reference article.

In order to increase the accuracy of the classification of unbalanced cancer data, the following suggestions can be used in future works:

1- Using fuzzy clustering to improve the performance of data balancing.

2- Using deep learning algorithms to improve classification performance.

3- Using feature selection algorithms to increase classification velocity and accuracy.

Acknowledgment

The research leading to these results has received no Research Project Grant Funding.

References

[1]. Aljabri, M. G., & Yousif, J. H. (2023). Blockchain Technology Effects on Healthcare Systems Using the IoT. In Intelligent Internet of Things for Smart Healthcare Systems (pp. 165-173). CRC Press.
[2]. Alam, T. M., et al. (2021). A machine learning approach for identification of malignant mesothelioma etiological factors in an imbalanced dataset. The Computer Journal, 64(5), 790–801.
[3]. Alkharabsheh, K., Alawadi, S., Kebande, V. R., Crespo, Y., Fernández-Delgado, M., & Taboada, J. A. (2022). A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of God class. Information and Software Technology, 143, 106736.
[4]. Almomany, A., Ayyad, W. R., & Jarrah, A. (2022). Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study. Journal of King Saud University-Computer and Information Sciences, 34(6), 3815-3827.
[5]. Dubey, S. (2022). A Comparative study of Breast Cancer Diagnosis and Classification Using Neural Networks and Machine learning models (Doctoral dissertation, Dublin, National College of Ireland).
[6]. Elwahsh, H., Tawfeek, M. A., Abd El-Aziz, A. A., Mahmood, M. A., Alsabaan, M., & El-shafeiy, E. (2023). A new approach for cancer prediction based on deep neural learning. Journal of King Saud University-Computer and Information Sciences, 35(6), 101565.
[7]. Engati. (2023). Engati. https://www.engati.com/glossary/imbalanced-dataset. Accessed [10 July, 2023].
[8]. Fotouhi, S., et al. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of Biomedical Informatics, 90, 103089.
[9]. Fujiwara, K., et al. (2020). Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. Frontiers in Public Health, 8, 178.
[10]. Gan, D., et al. (2020). Integrating TANBN with cost-sensitive classification algorithm for imbalanced data in medical diagnosis. Computers & Industrial Engineering, 140, 106266.
[11]. Han, X., et al. (2020). Deep doubly supervised transfer network for diagnosis of breast cancer with imbalanced ultrasound imaging modalities. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention.
[12]. Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.
[13]. Hesketh, R. (2023). Introduction to cancer biology. Cambridge University Press. Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. Diane Cerra.
[14]. He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.
[15]. He, H., & Shen, X. (2007). A Ranked Subspace Learning Method for Gene Expression Data Classification. In Proceedings of the International Conference on Artificial Intelligence.

[16]. Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2022). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences.
[17]. Jodha, R. (2023, June 12). KDD in Data Mining- Scaler Topics. Scaler Topics. https://www.scaler.com/topics/kdd-in-data-mining/ Accessed [10 July, 2023].
[18]. Kazem, H. A., Yousif, J. H., Chaichan, M. T., Al-Waeli, A. H., & Sopian, K. (2022). Long-term power forecasting using FRNN and PCA models for calculating output parameters in solar photovoltaic generation. Heliyon, 8(1).
[19]. KHAMIS, Y., & Yousif, J. H. (2022). Deep learning Feedforward Neural Network in predicting model of Environmental risk factors in the Sohar region. Artificial Intelligence & Robotics Development Journal, 201-2013.
[20]. Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232.
[21]. Lanjewar, M. G., Panchbhai, K. G., & Charanarur, P. (2023). Lung cancer detection from CT scans using modified DenseNet with feature selection methods and ML classifiers. Expert Systems with Applications, 224, 119961.
[22]. Liu, N., et al. (2020). A novel ensemble learning paradigm for medical diagnosis with imbalanced data. IEEE Access, 8, 171263-171280.
[23]. Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707.
[24]. López, V., et al. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141.
[25]. Miller, K. D., Nogueira, L., Devasia, T., Mariotto, A. B., Yabroff, K. R., Jemal, A., ... & Siegel, R. L. (2022). Cancer treatment and survivorship statistics, 2022. CA: a cancer journal for clinicians, 72(5), 409-436.
[26]. Pearson, R., Goney, G., & Shwaber, J. (2003). Imbalanced Clustering for Microarray Time-Series. In Proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Sets II.
[27]. Sarkar, P., & Srivastava, D. (2022, April). Computational Intelligence Approach to improve the Classification Accuracy of Brain Tumour Detection. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 406-414). IEEE.
[28]. Shen, C., Tannenbaum, D., Horn, R., Rogers, J., Eng, C., Zhou, S., ... & Dasari, A. (2022). Overall survival in phase 3 clinical trials and the Surveillance, Epidemiology, and End Results database in patients with metastatic colorectal cancer, 1986-2016: a Systematic Review. JAMA network open, 5(5), e2213588-e2213588.
[29]. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. Ca Cancer J Clin, 73(1), 17-48.
[30]. Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer statistics, 2021. Ca Cancer J Clin, 71(1), 7-33.
[31]. Sun, Y., Wong, A. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 687-719.
[32]. Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Addison Wesley.
[33]. Tang, X., et al. (2021). A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis with Transcriptome and Functional Proteomic. IEEE Access, 9, 51659-51668.
[34]. Thölke, P., Mantilla-Ramos, Y. J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., ... & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage, 277, 120253.
[35]. Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1), 6256.
[36]. Wang, Q., et al. (2020). Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis. Expert Systems with Applications, 152, 113334.
[37]. Weiss, G. (2004). Mining with rarity: a unifying framework. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, 6(1), 7-19.
[38]. Xie, C., et al. (2020). Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. European Journal of Nuclear Medicine and Molecular Imaging, 47(12), 2826-2835.
[39]. WHO. (2020). Cancer - Fact sheet. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed [10 July, 2023].
[40]. Yang, H., & Chen, Y.-P. P. (2015). Data mining in lung cancer pathologic staging diagnosis: correlation between clinical and pathology information. Expert Systems with Applications, 42(15), 6168–6176.
[41]. Yousif, J. (2019). Hidden Markov Model tagger for applications based Arabic text: A review. Journal of Computation and Applied Sciences IJOCAAS, 7(1).
[42]. Yousif, J. H. (2011). Information Technology Development. LAP LAMBERT Academic Publishing, Germany ISBN 9783844316704.
[43]. Yousif, J. H., Kazem, H. A., Al-Balushi, H., Abuhmaidan, K., & Al-Badi, R. (2022). Artificial Neural network modelling and experimental evaluation of dust and thermal energy impact on monocrystalline and polycrystalline photovoltaic modules. Energies, 15(11), 4138.