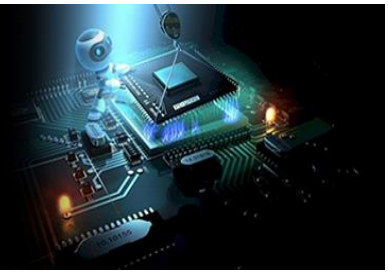
International Journal of Engineering in Computer Science



E-ISSN: 2663-3590
P-ISSN: 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2024; 6(1): 09-17
Received: 04-05-2024
Accepted: 13-06-2024

Christopher Francis Britto
Research Scholar, Department of Computer Science Mahatma Gandhi University Meghalaya, Meghalaya, India

Dr. Bidyut Kumar Das
Department of Computer Science, Mahatma Gandhi University, Meghalaya, Meghalaya, India

Yogesh V Patil
Assistant Professor, Department of Computer Application, Siddhant College of Management Studies, Pune, Maharashtra, India

Corresponding Author:
Christopher Francis Britto
Research Scholar, Department of Computer Science Mahatma Gandhi University Meghalaya, Meghalaya, India

Handling outliers and over fitting for lung cancer disease prediction through neural network clustering

Christopher Francis Britto, Dr. Bidyut Kumar Das and Yogesh V Patil

DOI: https://doi.org/10.33545/26633582.2024.v6.i2a.118

Abstract
With the rapid increase of cancer disease-affected people worldwide, predicting the cancer disease patients and diagnosing the cancer disease becomes crucial. The cancer disease diagnosis models have greatly confronted the data scarcity and often deal with the outliers. This paper presents the Neural network Clustering-based Outlier Handling (NCOH) methodology to improve the performance of multi-class cancer detection by modeling the outlier detection and outlier handling processes. The proposed methodology enhances the data augmentation process and the outlier handling process to enforce the cancer disease diagnosis using the oversampling and the neural network clustering methods. The NCOH approach handles the outliers with the impact of the clustering and the Z-score normalization score. The experimental results illustrate that the proposed approach significantly outperforms the existing outlier detection method by 15% of higher recall while testing the performance of the multi-class cancer classification model.

Keywords: Cancer diagnosis, over fitting, outlier, neural network clustering, autoencoder, K-Means clustering, SMOTE, and Z-score

1. Introduction
In the real world, cancer disease has increasingly impacted people by its different forms of cell growth. Early identification of the cancer disease in the human body is essential to provide the relevant diagnosis to the patients to decrease the death rate influenced by the cancer disease. Owing to the lack of adequate knowledge in the cancer patient samples, disease control becomes impossible over the diversified patterns of the cancer disease [1]. Scarcity of the medical data is the type of the classical issues for the design of the automatic classification system. Class imbalance is also one of the data scarcity constraints encountered by the classification model [2]. Detection of cancer diseases is challenging due to inadequate training knowledge of sensitive information or privacy constraint. The learning model necessitates the good representation of the original data to provide accurate decisions based on the underlying characteristics of the data. On the other hand, synthetic data creation is also influenced by the privacy reasons on the original data preservation. In particular, medical records have the most sensitive information about the patients; hence, generating sufficiently large data becomes crucial for assisting the clinical decision-making [3]. Most decision-making models confront the lack of knowledge in all the classes; hence, oversampling has been increasingly used as a viable solution to mitigate the scarcity of training or testing data. The synthetic Minority Oversampling Technique (SMOTE) [4] method is one of the widely used oversampling methods by the researchers, which generates the fixed number of synthetic instances based on the selected neighbors for every original instance through random interpolation. However, oversampling the input data without over fitting becomes challenging in the data augmentation or enrichment process.
Medical records comprise the outliers that significantly differ from the remaining observations, which provide potential insights into clinical decision-making. For instance, outliers in the ECG medical records represent the potential heart attacks over the normal physical conditions in the remaining observations of the patients. Outlier detection is the unsupervised model due to the lack of prior labels for the outliers in the different domains. Most of the existing outlier detection methods [5] have applied statistical models and clustering methods. Recently, several researchers have adopted neural network-based

autoencoders to detect the outliers through the inherent feature analysis [6, 7]. In essence, reconstructing the input data from the compact and hidden representation of the data tends to avoid outliers due to the only consideration of the representative features in the input. However, the reconstruction error-based outlier detection in the input data is likely to mislead the medical diagnosis by compromising the abnormal observations with the normal observations in the medical records [8]. Hence, instead of detecting the outliers and normalizing the outliers without the risk analysis [9], contextually handling the outliers is significant to leverage the accurate diagnosis of the cancer disease [10]. Thus, this work focuses on addressing the data scarcity and the outlier handling constraints to facilitate the cancer disease diagnosis. The proposed methodology employs the SMOTE oversampling method and the autoencoder neural network with the K-means clustering method to accomplish these objectives. In essence, the proposed approach enhances the oversampling with the optimal neighbor selection for the data augmentation and the neural network clustering with the statistical outlier detection methods for the outlier detection and handling. Thus, the NCOH approach effectively avoids the over fitting and misclassification or fault diagnosis of the cancer disease.

The main contributions of the proposed Neural network Clustering-based Outlier Handling (NCOH) methodology are presented as follows.

- This work presents the model for the data augmentation and the outlier handling for the accurate cancer disease diagnosis with the help of the SMOTE oversampling and the neural network clustering methods.
- The NCOH approach contextually enriches the cancer patient samples through optimal neighbor selection based SMOTE method when there is an inadequate amount of the cancer patient samples, which also avoids the over fitting.
- Applying the neural network clustering involving the auto encoder and the K-Means clustering inherently analyzes the patient behaviors and groups the similar risk patients in an unsupervised manner. In subsequence, it employs the statistical methods of the IQR and the Z-score to detect the outliers that exist in the symptom-specific clusters.
- Also, the NCOH approach computes the score for the outlier data points by utilizing the attribute-values of the patient samples in the perspective of the corresponding clusters along with the Z-score to assist the risk level of the cancer disease classification, which mitigates the fault diagnosis.
- Thus, the experimental model tests the impact of the handled outlier in the cancer disease classification process while implementing the cancer patient samples with the lung cancer dataset.

Related Works

This section reviews several existing research works addressing data augmentation, over fitting, outlier detection and handling, and accurate cancer disease detection constraints.

Data Enrichment and Over fitting Handling Approaches

Recent approach [11] employs the Convolutional Neural Network (CNN) to classify the histological routine colon cancer nuclei. By modeling 1, 512, 868 learnable parameters in the CNN, it mitigates the over fitting in the classification results while testing the CRCHistoPhenotypes dataset. Deep Adversarial Data Augmentation (DADA) approach [12] addresses the data augmentation problem in training the supervised and class conditional Generative Adversarial Network (GAN) model through the discriminator loss-based data augmentation. It computes the discriminator loss and enforces the real and augmented samples to determine the effective decision boundaries. The research work [13] generates the training data for the imbalanced dataset using the GAN model, which is similar to the ADASYN or SMOTE method processes. It is demonstrated that the performance of the generated dataset is high compared to the original dataset. The consistent Risk Correction model [14] handles the over fitting constraint with the help of the Unlabeled-Unlabeled (UU) method, which analyzes the negative empirical risk and wraps the terms based on the correction functions. Deriving the estimation error bound minimizes and maintains the consistency of the corrected risks. The research work [15] applies non-parametric learning and analyzes the multivariate dataset, which leverages the generation of new synthetic samples. It improves the generalization capabilities of the learning model and reduces the outliers in the generated dataset with the assistance of the autoencoder model and Restricted Boltzmann Machine. The automatic Relational Data Augmentation (ARDA) approach [16] utilizes the data repository and input dataset to generate the augmented data based on two distinct components. It examines the different attributes of the input data to join the data with the input data and applies the Random Injection Feature Selection (RIFS) algorithm to remove the irrelevant features from the result of the join data. The research work [17] presents an adaptive discriminator augmentation model to maintain the training data with the limited regimes data for the GAN model. By augmenting the data for training the GAN, it improves the classification performance without considering the dataset properties and the training data size. Correlation Coefficient-Aware Data Augmentation model [18] improves the performance of the machine learning model by extracting the prominent features from the massive set of features. By computing the non-linear correlation based on the new index structure, it enriches the data based on the rank-based correlation of the data. The generative Adversarial Network Synthesis for Oversampling (GANSO) approach [19] applies both the GAN and markov random fields to handle the scarcity constraint of the training data. It also measures the clique similarity based on the analysis of the linear discriminant on the features, which enforces the classification of the functional Magnetic Resonance Imaging (fMRI) data, sleep stages of Electroencephalographic (EEG), and Electrocardiographic (ECG) data. The research work [20] presents the hybrid sampling method involving the Synthetic Minority Oversampling Technique (SMOTE) and the Weighted Edited Nearest Neighbor Rule (WENN). It applies the k-Nearest Neighbor (kNN) method and the weighted distance function to detect and ignore the majority and minority class samples.

Outlier Detection and Cancer Prediction Approaches

The outlier detection model [21] employs the deep-stacked auto encoder and density-based clustering to detect the deviations in the behavioral patterns of the vessels.

Applying the unsupervised learning model of the stacked auto encoder maps the track history as the significant feature vectors for vessels and, thus, determines the possible outliers from the clustering results. Single-Objective Generative Adversarial Active Learning (SO-GAAL) approach [22] performs the mini-max game between the generator and discriminator to detect the outliers. Determining the stop node of training through an unsupervised model provides adequate learning information in the accurate representation for the entire dataset. The research work [23] presents the Clustering with Outlier (COR) algorithm considering the joint cluster analysis to detect the outliers. It applies the K-means clustering and enhances the clustering compactness based on the Holo entropy, which leverages the removal of the outliers through the generation of the basic partitions. Time series outlier detection model [24] employs the recurrent autoencoder ensembles with the help of the Sparsely-connected Recurrent Neural Networks (S-RNNs). It combines the independent framework and shared framework associated with the multiple S-RNN based autoencoders to detect the outliers and mitigate the outliers' over fitting effects. The research work [25] detects the outliers based on the reconstruction errors and the identification criterion of the Grubbs and PauTa criteria. Analyzing larger reconstruction errors from the stacked autoencoder results distinguishes the characteristics between the abnormal data and the normal data and locates the outliers in the input data effectively. The research work [26] introduces the algorithmic variations and augments the Dispersion Entropy (DisEn) for the physiological signal monitoring applications. By mitigating the disruptive effect of the outlier samples, it improves the performance of the entropy quantification algorithms in the time series application. Cervical Cancer Prediction Model (CCPM) [27] analyzes the risk factors of cervical cancer disease and removes the outliers through isolation forest (iForest) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms. Moreover, it increases the number of samples using the SMOTE and SMOTE with Tomek link (SMOTETomek) models to detect the outliers in the cervical cancer disease data and thus, improves the performance of the cervical cancer disease prediction.

Problem Statement

The disease prediction models often confront inadequate training knowledge about the risk levels in the medical field. Even though several data augmentation methods have enriched the small-scale data to the large-scale data, it is likely to comprise the noisy data in the enriched patient samples. In the prediction model, over fitting affects the decision-making accuracy when augmenting the data with a new set of medical observations. In addition, medical data observations contain outliers due to the continuous variations in the health conditions of the patients. Applying statistical methods for outlier detection leads to inaccurate decision-making over the multiple levels of cancer risk patients. Most conventional systems detect and ignore the outliers, which misguides the accuracy of the disease diagnosis. Handling the outlier data using the entire cancer patient samples leads to inaccurate score computation when there is the existence of different cancer patient risks. Also, data mining and machine learning models based on outlier detection or pattern extraction models have confronted the lack of inherent feature analysis among the different categories of cancer patients. Although, several unsupervised learning models fail to determine and handle the outliers among the multiple cancer risk patients. Hence, it is essential to detect and handle the outliers without over fitting and improved cancer disease diagnosis.

The Proposed NCOH Methodology

With the target of accurately diagnosing the cancer patients, the NCOH approach focuses on resolving the over fitting and outlier constraints through the neural network clustering method. To support the neural network-assisted clustering for small data samples, the NCOH approach applies the SMOTE oversampling with the optimal patient set selection while generating the new samples. Applying the autoencoder and the K-means clustering methods facilitates the detection of the outliers when determining the outlier range based on the IQR and the Z-score methods. Furthermore, the NCOH approach handles the outliers based on the knowledge of the clustered patients and the Z-score without compromising the outlier values with the different levels of the cancer patients. Figure 1 illustrates the overall process of the outlier detection and handling methodology for the multi-class cancer classification.

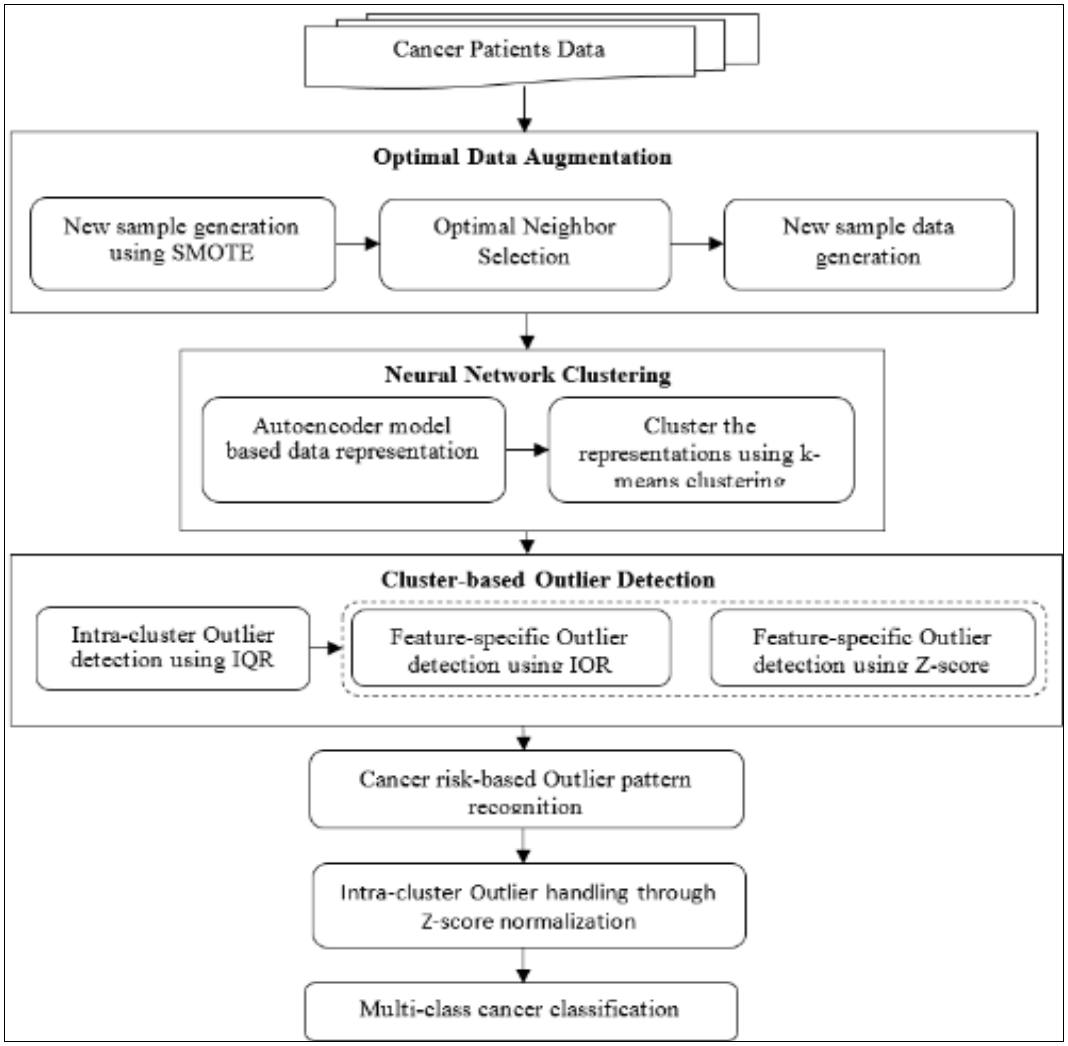


Fig 1: Illustration of the overall process of the outlier detection and handling methodology for the multi-class cancer classification

Optimal Data Augmentation through Oversampling

The NCOH approach increases the input knowledge from the available class-specific data through the oversampling method to deal with the data scarcity in the medical domain, especially cancer patients' data. The NCOH approach employs the Synthetic Minority Oversampling Technique (SMOTE) to enrich small cancer samples. With the target of alleviating the noise in the generated new samples, it adds the new samples in the input dataset for each class and considers the significant attributes. The SMOTE oversampling method has been increasingly used to handle the high imbalance data over the majority and minority classes through the balanced distribution of the samples. Consequently, it improves the classification accuracy of the neural network model with the knowledge of the large-scale input data, which avoids the over fitting in the cancer detection or classification results. The cancer patient dataset comprises 'N' patients, Pi={X1, X2, XN}, and each patient (Pi) contain 'M' attributes or symptoms, A={Y1, Y2, …, YM}. The NCOH approach optimally generates the new samples or cancer patient records based on the effective neighbor selection in the input patients.

$$N(S_S)=\begin{cases} C_S \subseteq |S| \\ G_S \subseteq |C_S| \\ A_S \subseteq |G_S \subseteq C_S| \end{cases} \quad (1)$$

The NCOH approach initially prefers the non-medical attributes or features such as the 'Age' and 'Gender' under the 'class' or 'level' attribute to increasing the dataset size with the assistance of the SMOTE oversampling method. SMOTE method generates artificial samples based on the nearest neighbors in the minority class to create the balanced distribution of the samples over the multiples without duplications and noises. Instead of randomly selecting the nearest neighbors, the NCOH approach contextually selects the nearest neighbors to facilitate the sensitive cancer disease diagnosis over the multiple classes. Equation (1) describes the neighbor selection criteria (N) in the proposed SMOTE process for reducing the over fitting during the augmentation of the scarce cancer patient records. 'SS' denotes the subset of the samples as the neighbors for the oversampling. In the SMOTE model, the neighbor selection depends on three features such as the class-specific samples (CS), gender-specific samples (GS), and age-specific samples (AS) in the overall cancer, patient samples (S). The NCOH approach targets selecting the subset of samples for generating the related neighborhood data through SMOTE oversampling. The hierarchy level attribute consideration enforces the accurate pattern generation for each class without the noisy data from the perspective of the top-level attributes of 'Gender' and 'Age'. In the proposed methodology, the SMOTE method computes the distance (d) between each patient sample (Pi) in the overall patients (S) and its nearest neighbor (Pj) from

its class ({CS(i)}), gender ({GS(i)}), and age ({AS(i)})-related set of patients. The proposed data augmentation model optimally selects the neighbors from the hierarchically filtered neighbor patients. Thus, the NCOH approach generates the new samples contextually with the target of avoiding the over fitting and cancer disease decision-making performance over the small-scale patient samples.

Neural Network Clustering Based Outlier Detection

With the target of facilitating the accurate cancer diagnosis, the NCOH approach focuses on detecting the outliers in the health behaviors of the cancer patients. To accomplish this, the NCOH approach applies the unsupervised neural network model that inherently analyzes the representative patterns of the cancer patients and assists the detection of the outliers within a particular category or group of the patients. It employs the autoencoder as the neural network model, which learns the data in an unsupervised manner and reduces the gap between the input and output data. The autoencoder model compresses the large-scale input data to a small dimension and reconstructs the output from the small dimension. In subsequence, it updates the weight of the learning model based on the difference between the input data and the reconstructed data values. The reconstruction of the data is likely to compromise the outlier data with low-dimensional projection, which results in high reconstruction error. The autoencoder model plays a significant role in outlier detection by extracting and learning the distribution characteristics from the input data through its deep structure. In the proposed methodology, the autoencoder model can tune the learning parameters to preserve the potential information in the input data without the information loss, which leads to maintaining the optimal reconstruction error. It generates the representative patterns of the cancer patients with components of the encoder and decoder, and the re-constructor.

$$\min_{i \in N} E(i,i') = \begin{cases} H\left(B(i) \middle| B(i')\right), & \text{if } i \in \{0,1\} \\ \|i-i'\|^2, & \text{if } i \in R \end{cases} \quad (2)$$

Equation (2) estimates the capability of the reconstruction in the autoencoder model that is minimal reconstruction error, which is accomplished by the parameter optimization of the learning model. It computes the error between the vectors of input (i) and the vectors of reconstructed input (i') based on the Bernoulli cross-entropy. B(i) and B(i') refer to the mean values of the input and reconstructed input, respectively. By applying the autoencoder neural network model, the NCOH approach represents the influential patterns for cancer patients at different levels of cancer risks. In subsequence, the NCOH approach targets detecting the outliers by utilizing the results obtained from the neural network and the help of the clustering method. To accomplish this, the NCOH approach applies the K-means clustering to group the cancer patients into multiple clusters. Clustering plays a significant role in recognizing the outlier data points in the feature-specific grouping, symptoms, or human behaviors in the cancer patient records, which becomes effective when grouping the entire patterns of the patients. In some scenarios, the clustering method compromises the outliers with the normal data points when there are different categories, such as risk levels of the cancer patients. For instance, if the 'low' risk patient has an outlier value in any one of the symptoms, the clustering model groups patients with either 'medium' or 'high' risk patient clusters. Hence, determining the outliers within the corresponding patient groups is a challenging task. To address this constraint, the NCOH approach clusters the patients based on the analysis of the entire patterns using the K-means clustering method and then utilizes the clustering results for further process of the outlier detection.

In the proposed methodology, the outlier detection process heavily relies on the Inter Quartile Range (IQR) and Z-score methods for the clustered patients, which assists to precisely detect the uncertain values or outliers. The NCOH approach applies the IQR method to detect the outliers that exist in the features or symptoms of the clustered patients. Also, it applies the Z-score method to detect the outliers in the patterns of the clustered patients. In consequence of the outlier detection methods, the NCOH approach determines the final outlier in the clustered patients based on the conditions mentioned in Equation (3). From the results of the IQR and Z-score methods, the NCOH approach stores a set of patients with outlier values ((Oi)IQR) and a set of patients with outlier values ((Oi)Z-score) respectively. If both outlier sets comprise the same patients, the NCOH approach decides that the patient has outliers in their attribute values.

$$O_i^C = \begin{cases} 1, \text{if } (O_i)_{IQR} = (O_i)_{Z\text{-score}} \\ 0, \text{if } (O_i)_{IQR} \neq (O_i)_{Z\text{-score}} \end{cases} \quad \text{for } i \in n \quad (3)$$

By jointly considering two different outlier detection methods, the NCOH approach detects the outlier values for the patients in each cluster from the perspective of the features or symptoms. In Equation (3), 'n' indicates the total number of patients in each cluster, n "⊆" N. 'n' varies for the cluster that is each cluster has a different number of cancer patients with various outlier values in their symptoms. Thus, the NCOH approach effectively determines the outliers in the cancer patient records with the help of neural network clustering and the hybrid outlier handling methods such as the IQR and the Z-score methods.

Cancer Risk-based Outlier Handling

In the subsequence of detecting the outliers, the NCOH approach focuses on handling the outliers in the cancer patient records rather than removing the outliers from the input data. Ignoring the outliers leads to inaccurate decision-making, particularly because the outliers in the cancer patient behaviors leverage the fault diagnosis. Hence, the NCOH approach analyzes the patient behaviors in the clusters to handle the outliers determined in the corresponding clusters based on the cancer risk levels. In essence, the NCOH approach normalizes the outliers based on the Z-normalization score and assigns the weight to the outliers based on neighbor patients in a particular cluster. The knowledge of the clustered patients greatly assists the handling of the outlier data points to improve the cancer disease diagnosis accuracy. By applying Equation (4), the NCOH approach computes the normalized value for the detected outliers in the clusters.

$$\begin{cases} \frac{\left(\left[\alpha\times V(O_i^C)\right]+\left[\beta\times V_Z(O_i^C)\right]\right)}{2}+\gamma\times\left(\sum_{i=1}^{n_C-1}\frac{V(P_i^C)}{n}\right), & \text{if } V(O_i^C)>\frac{V(P_i^C)}{n} \\ \frac{\left(\left[\alpha\times V(O_i^C)\right]+\left[\beta\times V_Z(O_i^C)\right]\right)}{2}-\gamma\times\left(\sum_{i=1}^{n_C-1}\frac{V(P_i^C)}{n}\right), & \text{if } V(O_i^C)<\frac{V(P_i^C)}{n} \end{cases}$$

In equation (4), "α, β, and γ" represent the constant weight parameters of 0.125, 0.25, and 0.625, respectively, to compute the normalized outlier values. ⟦"V(O" ⟧_"i" ^"C" ") and " ⟦"V" _"Z" "(O" ⟧_"i" ^"C" ")" denotes the outlier value in the original data and the outlier value after the Z-score normalization, respectively. ⟦"V(P" ⟧_"i" ^"C") indicates the attribute value of other patients in a particular attribute or symptoms in the corresponding cluster. Suppose the detected outlier value is greater than the average of other patients' attribute value. In that case, the NCOH approach handles the outlier value based on adding the outlier value with the attribute values of other patients in a particular cluster. In other cases, it subtracts the outlier value with the attribute-values of other patients in a particular cluster. The NCOH approach considers only the clustered patients for the outlier data handling rather than considering the diversified patient behaviors. Thus, the NCOH approach effectively handles the outlier values in the context of the cancer risk or category-based symptoms analysis to improve the cancer diagnosis performance.

Experimental Evaluation

To assess the performance of the outlier detection model, the experimental model evaluates the proposed methodology with the existing IQR and Mean value-based Outlier Handling (IMOH) model. The comparative outlier detection and handling model refers to the IQR-based outlier detection and mean value-based handled outliers while testing with the machine learning model.

Experimental Setup

The experimental framework implements the proposed outlier detection and handling model on Ubuntu 16.04 64-bit machine with a 3GHz Intel CPU and 16GB memory. It illustrates the outlier detection and handling algorithm on the multi-class lung cancer detection using the Python programming language. Moreover, to conduct the experiments of the neural network model, the experimental framework employs the python libraries with the python version 3.6.8. The experimental model employs the cancer disease dataset that consists of diversified cancer risk patients with different symptoms to implement the proposed outlier detection and handling algorithm. To test the impact of the outlier handling on the cancer classification model, the experimental model performs the classification of the patient records using a decision tree machine learning algorithm. The NCOH approach employs the Lung cancer dataset [28] to test the performance improvement of the NCOH approach, which consists of 1000 patient records and 24 attributes or features with the label. In the lung cancer dataset, the attributes namely Patient ID, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing, Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, and Snoring. In the multi-class lung cancer dataset, the labels are categorized into Low, Medium, and High. The experimental model creates the outlier data in the extracted lung cancer dataset to demonstrate the outlier detection and outlier handling processes.

Evaluation Metrics

To validate the effectiveness of the proposed methodology, the experimental framework employs the following evaluation metrics.

Precision: It is the ratio between the numbers of correctly identified level of cancer disease samples to the total number of cancer disease samples in the different levels recognized by the system.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall or True Positive Rate: It is the ratio between the number of correctly identified levels of cancer disease samples to the total number of disease instances that actually belong to the level of cancer disease.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

F-Measure: It is the harmonic mean of Precision and Recall.

$$\text{F-Measure} = \frac{2\times(\text{Precision}\times\text{Recall})}{(\text{Precision}+\text{Recall})}$$

G-Mean: It is the geometric mean of the classification performance, which has been effectively used to measure the performance of the classification over the different percentages of majority and minority samples.

$$\text{G-mean} = \sqrt{\text{True Positive Rate}\times\text{True Negative Rate}}$$

$$\text{True Negative Rate} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

Experimental Results

This section compares the NCOH approach with the existing approach for the four different performance metrics while varying the number of patient samples.

Number of Samples Vs. Precision

Figure 2 illustrates the precision performance of the proposed NCOH, and the existing IMOH approaches for the different number of patient samples from 6000 to 30000 with the interval of 6000 samples. The proposed NCOH approach yields 9.33% higher precision than the existing IMOH approach when the number of samples is 30000 with the possible outliers. The NCOH approach applies neural network clustering to group similar risk patients to improve the outlier handling performance within the cluster. In contrast, the existing model handles the outliers by replacing the detected outlier value with the mean of the entire patient. Thus, the existing IMOH approach achieves the reduced performance of the multi-class cancer classification due to the lack of contextual or similar patient analysis.

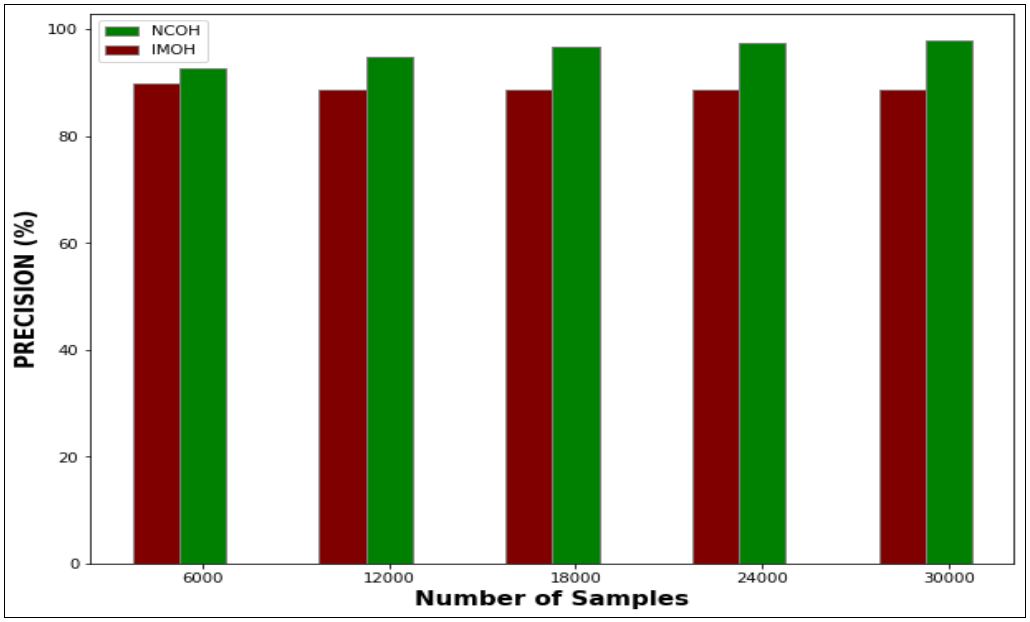


Fig 2: Number of Samples Vs Precision

Number of Samples Vs. Recall

The recall performance of both the proposed and existing outlier handling approaches while varying the number of samples from 6000 to 30000 is depicted in Figure 3. The proposed NCOH approach linearly increases the recall performance by 4.817% compared to the existing IMOH approach, even when the number of samples is a minimum of 6000. The existing IMOH approach fails to focus on the importance of the outlier values on the class labels, which leads to the misclassification of the cancer risk patients. At the same time, the proposed NCOH approach computes the normalized outlier value based on the mean value of the patients in that particular cluster, outlier value, and Z-score-based outlier value. As a result, the proposed approach maintains the significance of the outliers during the classification of the multi-class cancer diseased patients. Moreover, by effectively oversampling the patient samples, it avoids over fitting in the multi-class cancer classification, which increases the effectiveness of the proposed NCOH approach.

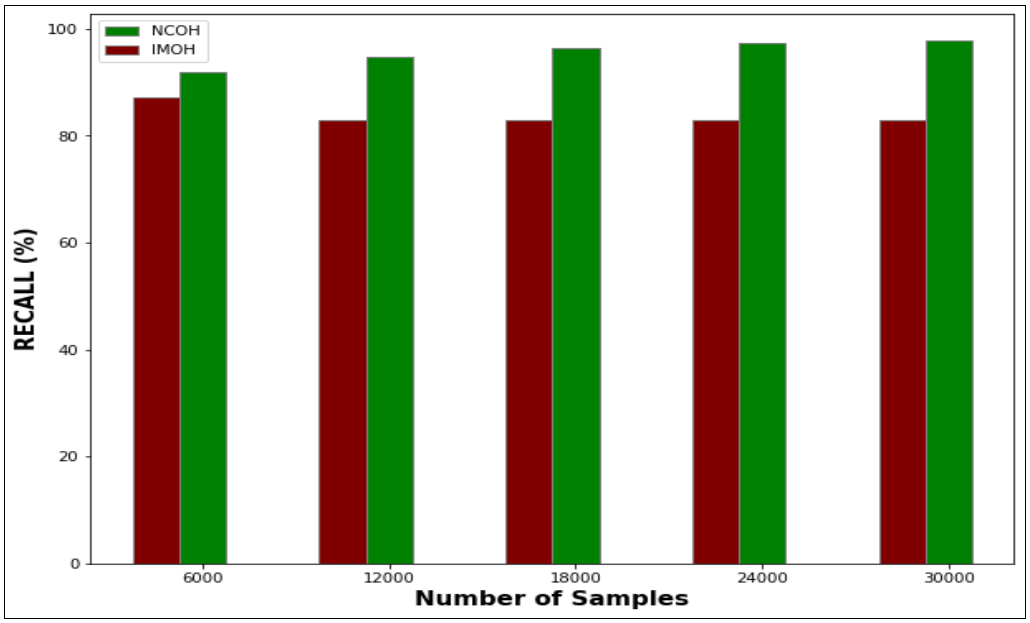


Fig 3: Number of Samples Vs Recall

Number of Samples Vs. F-Measure

Figure 4 shows the F-Measure of the proposed NCOH compared to the existing IMOH approach with the increase of the number of samples from 6000 to 30000. The proposed NCOH approach obtains a higher F-Measure value even when there is a lack of available patient records by applying the SMOTE oversampling, which enriches the size of the lung cancer dataset. Moreover, it applies the autoencoder-based neural network clustering for effectively clustering the patients based on the inherent feature analysis. As a result, the proposed NCOH approach yields 1.3% of higher F-Measure than the existing IMOH approach when varying the number of samples from 18000 to 30000. In addition, mean value-based outlier handling in the existing IMOH approach compromises the outliers and misleads the inaccurate cancer diagnosis.

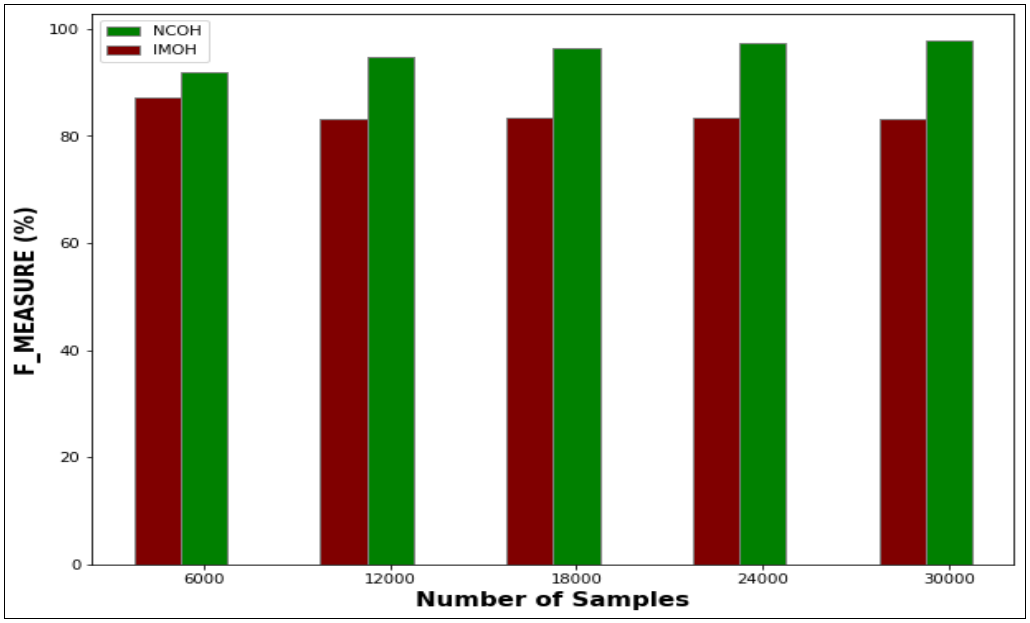


Fig 4: Number of Samples Vs F-Measure

Number of Samples Vs. G-Means

G-mean of the proposed NCOH approach and the existing IMOH approach for the different number of patient samples from 6000 to 30000 are illustrated in Figure 5. The performance of the G-mean increases with the increase of the number of patient records by utilizing the knowledge of patient records. The proposed NCOH approach obtains a 98.394% G-mean value when the number of samples is 30000. In the same scenario, the existing IMOH approach only 86.106% of G-mean value due to the misclassification of the cancer risk patients by the mean value replacement for the outlier values among the low, medium, and high-risk cancer patients. As a result, the proposed NCOH approach accomplishes 3.972% higher G-mean performance than the existing IMOH approach, even when there is minimal knowledge of cancer patients.

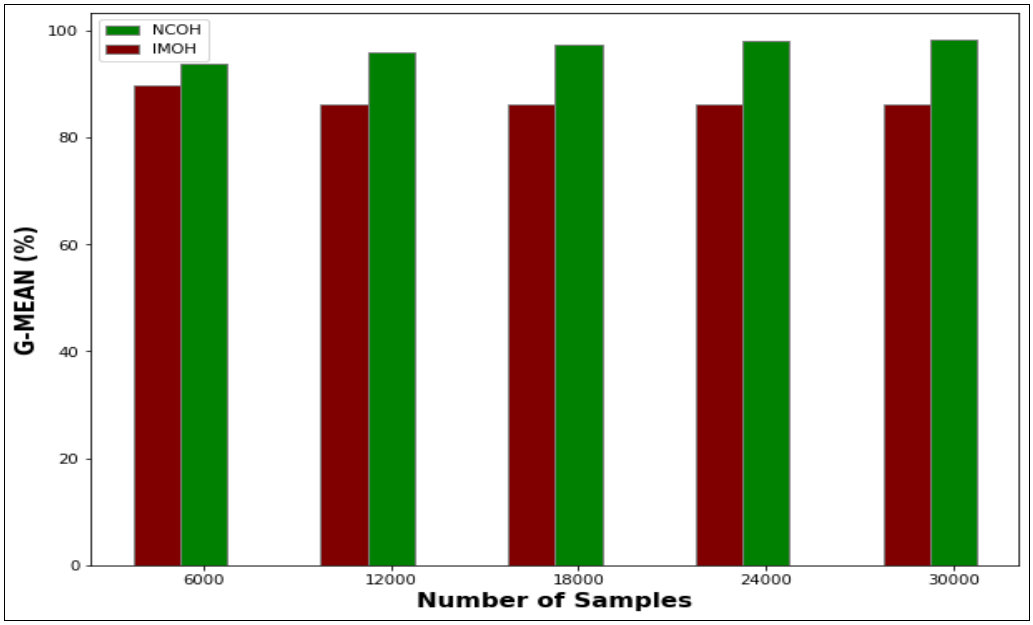


Fig 5: Number of Samples Vs G-Mean

Conclusion

This work has presented the neural network clustering-based cancer disease outlier detection and handling methodology with statistical methods such as the IQR and Z-score normalization. It has resolved the over fitting constraint and the data scarcity in the multi-class classification through the optimal augmentation of the patient samples using the SMOTE model. By applying the unsupervised neural network model and the K-means clustering, the NCOH approach has facilitated the detection of the outliers in the data-sensitive cancer patient samples and the help of the IQR and Z-score normalization. From the results of the neural network clustering, the NCOH approach effectively handles the outliers in the cancer patient samples and leverages the accurate multi-class classification of cancer patients without over fitting. Thus, the NCOH approach improves the cancer disease classification performance by 12.28% higher G-mean than the baseline model, even when there are outliers in a large number of input samples.

References

1. Kharya S. Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint arXiv:1205.1923; c2012.
2. Wen Q, Sun L, Yang F, Song X, Gao J, Wang X, *et al*. Time series data augmentation for deep learning: A survey. arXiv preprint arXiv:2002.12478; c2020.
3. Santoso B, Wijayanto H, Notodiputro K, Sartono B. Synthetic over sampling methods for handling class imbalanced problems: A review. In: IOP conference series: earth and environmental science. Vol.58, No.1. p.012031. IOP Publishing; c2017.
4. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
5. Wang H, Bah M, Hammad M. Progress in outlier detection techniques: A survey. IEEE Access. 2019;7:107964-108000.
6. Xia Y, Cao X, Wen F, Hua G, Sun J. Learning discriminative reconstructions for unsupervised outlier removal. In: Proceedings. of IEEE International Conference on Computer Vision. 2015:1511-1519.
7. Kieu T, Yang B, Jensen C. Outlier detection for multidimensional time series using deep neural networks. In: 2018 19th IEEE International Conference on Mobile Data Management (MDM). 2018:125-134.
8. Domingues R, Filippone M, Michiardi P, Zouaoui J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition. 2018;74:406-421.
9. Boukerche, Zheng L, Alfandi O. Outlier Detection: Methods, Models, and Classification. ACM Computing Surveys (CSUR). 2020;53(3):1-37.
10. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9(01).
11. Basha S, Ghosh S, Kishan Babu K, Ram Dubey S, Pulabaigari V, *et al*. Rccnet: An efficient convolutional neural network for histological routine colon cancer nuclei classification. In: IEEE 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV). 2018:1222-1227.
12. Zhang X, Wang Z, Liu D, Ling Q. Dada: Deep adversarial data augmentation for extremely low data regime classification. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019:2807-2811.
13. Tanaka F, Aranha C. Data augmentation using GANs. arXiv preprint arXiv:1904.09135. 2019.
14. Lu N, Zhang T, Niu G, Sugiyama M. Mitigating over fitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In: International Conference on Artificial Intelligence and Statistics. 2020:1115-1125.
15. Kondratyev, Schwarz C, Horvath B. Data anonymisation, outlier detection and fighting over fitting with Restricted Boltzmann Machines (January 27, 2020). 2020.
16. Chepurko N, Marcus R, Zgraggen E, Fernandez R, Kraska T, Karger D. ARDA: automatic relational data augmentation for machine learning. arXiv preprint arXiv:2003.09758. 2020.
17. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676. 2020.
18. Esmailoghli M, Quiane-Ruiz J, Abedjan Z. COCOA: COrrelation COefficient-Aware Data Augmentation. EDBT. 2021.
19. Salazar, Vergara L, Safont G. Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets. Expert Syst Appl. 2021;163:113819.
20. Guan H, Zhang Y, Xian M, Cheng H, Tang X. SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling. Appl Intell. 2021;51(3):1394-1409.
21. Protopapadakis E, Voulodimos A, Doulamis A, Doulamis N, Dres D, Bimpas M. Stacked autoencoders for outlier detection in over-the-horizon radar signals. Comput Intell Neurosci. 2017;2017.
22. Liu Y, Li Z, Zhou C, Jiang Y, Sun J, Wang M, *et al*. Generative adversarial active learning for unsupervised outlier detection. IEEE Trans Knowl Data Eng. 2019;32(8):1517-1528.
23. Liu H, Li J, Wu Y, Fu Y. Clustering with outlier removal. IEEE Trans Knowl Data Eng. 2019.
24. Kieu T, Yang B, Guo C, Jensen C. Outlier Detection for Time Series with Recurrent Autoencoder Ensembles. In: IJCAI. 2019:2725-2732.
25. Wan F, Guo G, Zhang C, Guo Q, Liu J. Outlier Detection for Monitoring Data Using Stacked Autoencoder. IEEE Access. 2019;7:173827-173837.
26. Kafantaris E, Piper I, Lo T, Escudero J. Augmentation of dispersion entropy for handling missing and outlier samples in physiological signal monitoring. Entropy. 2020;22(3):319.
27. Ijaz M, Attique M, Son Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors. 2020;20(10):2809.
28. Lung Cancer Dataset, Available Online at: https://data.world/cancerdatahp/lung-cancer-data, Accessed On May, 2021.