International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST52411130

Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence

Arati K Kale *1, Dr. Dev Ras Pandey2

*1Department of Computer Science and Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India
2Department of Computer Science and Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India

ARTICLEINFO

Article History:

Accepted: 15 Jan 2024
Published: 29 Jan 2024

Publication Issue :
Volume 11, Issue 1
January-February-2024
Page Number :
299-309

ABSTRACT

Healthcare datasets frequently contain large dimensional, distorted, uneven, missing, and imbalanced data. These difficulties may lower the effectiveness of machine learning algorithms. Before using machine learning algorithms for healthcare datasets, pre-processing is necessary to ensure the data is adequate for learning. The data pre-processing is essential to improve the performance of classification or prediction. This paper proposes a data pre-processing technique for enhancing healthcare data quality using artificial intelligence. The pre-processing includes handling missing values, outlier detection and handling imbalanced data. The missing values are imputed using the KNN-based approach, the outliers are detected using a cluster-based algorithm, and SMOTE and the Random resampling approach can rebalance the imbalanced data. Different machine learning classification algorithms are used to analyze the data quality. The real-time healthcare dataset is used to evaluate the performance of the proposed approach using accuracy, sensitivity, specificity, precision and f-measure. This research shows that the pre-processing techniques chosen have a considerable positive impact on the model's performance when comparing the model's efficiency with and without pre-processed data.

Keywords : Data Pre-Processing, Healthcare Data, Artificial Intelligence, Data Quality, Medical

I. INTRODUCTION

Contemporary healthcare information systems use medical databases to hold vast amounts of data. It promotes retrieving practical information from medical databases, offering insightful information for medical decision assistance [1]. Medical data classification aims to raise the high level of healthcare by developing classification models using medical datasets. The prediction and diagnostics can be made with medical data classification. The quality of the data affects the forecasts' classification accuracy.

Missing values, noise, inconsistent behaviour, excessive data, and class imbalances are all possible in data generated from several sources. This incomplete data needs to be cleaned up and prepared for further study at the data preparation stage [2].

Data pre-processing is converting a raw dataset into a valuable and intelligible dataset to enhance the performance of machine learning tasks clustering, classification, prediction, etc.,[3]. Data pre-processing includes data cleaning, normalization [4], discretization, missing value imputation [5], feature selection [6], outlier identification [7], and handling imbalanced data [8]. This paper considers the missing values, outliers and imbalanced data problems in medical data.

Clinical research frequently deals with missing data. Missing data happens when some or all of the sample's individuals do not have their values for the parameters of interest measured or recorded. There are several reasons why data may be absent [9], such as (i) patients refusing to answer specific questions, (ii) patients being lost to follow-up, (iii) physical or investigator error, and (iv) medical professionals not requesting specific tests for particular patients. Missing data are typically handled inadequately in the clinical research paradigm [10]. The most popular method is the complete case analysis, which removes rows that have missing values in the outcome or predictor variables. This decision is troublesome because it results in a model with poor generalization capabilities and a smaller dataset.

Furthermore, this approach often yields findings and mistakes that are optimistic in reality but may be minimal for the entire subset of data. Again, various studies may employ distinct subsets of the same dataset; for example, rows may be omitted in place of columns, or both may be eliminated; this decision complicates comparisons. Simple imputation using the mean and mode technique [11] is another method for resolving this problem. Although they produce a complete dataset, they are overly simplistic and so attribute values that are not realistic. This paper uses the K-Nearest Neighbor algorithm-based approach to impute the missing value.

One of the most essential data pre-processing techniques is outlier detection. Typically, anomalous objects in a dataset deviating from the general model are called outliers. Either noise or significant information can be found in outliers. The standard techniques for identifying outliers are density, statistics, distance, and grouping-based approaches [12]. Data mining's outlier identification is a significant subfield with extensive application across multiple sectors. In the medical field [13], outlier detection is typically used to track the pattern of diagnostic data and forecast disease outbreaks based on the presence of outliers. This paper uses a cluster-based approach to detect outliers in medical data.

Class imbalance appears difficult for many researchers in all data domains, regardless of application. When there are more instances of one class (majority) than of another (minority), this is referred to as a class imbalance [14]. A model's ability to predict the future may be hampered by learning datasets with this problem. It leads to a more biased model favouring the majority class and raising the misclassification rate. Class imbalance problems can be addressed using various techniques and broadly categorized into the following groups [15]: cluster-based approaches, algorithm-level approaches, data-level approaches, hybrid approaches and ensemble learning-based approaches. This paper uses a data-level and cluster-based approach to handle imbalanced data problems.

This paper proposes a data pre-processing technique for enhancing healthcare data quality using artificial intelligence. The pre-processing includes handling missing values, outlier detection and handling imbalanced data. The key objectives of this research paper are:

- This work presents a k-nearest neighbor-based algorithm to impute the missing values.
- The cluster-based algorithm is applied to detect the outliers in medical data.

- The SMOTE (Synthetic Minority Over-Sampling Technique) and random resampling approach is suggested to rebalance the imbalanced data.
- The proposed approach is evaluated using real-time medical data.

The remaining part of this research paper is organized as follows: Section 2 describes the related work, including missing value estimation, outlier detection and handling imbalanced data. Section 3 explains the proposed pre-processing methodology. The result and discussion are presented in section 4, and section 5 concludes the research paper.

II. RELATED WORK

2.1 Missing Value Estimation

Extreme gradient boosting (XGBoost), a supervised machine learning technique, is combined with an unsupervised prefilling approach by Zhang et al. [16] to forecast the missing values. The SMILES model is presented to restore the missing values of laboratory test variables. Prefilling, window-size-based feature extraction, and machine learning are all integrated into the model. The prefilling approach and the supervised method, which concurrently utilizes the longitudinal and cross-sectional context, can significantly enhance missing data imputation on the temporal variables.

A novel approach to imputation for symbolic regression with missing data was presented by Al-Helali et al. [17]. This method will impute missing variables for symbolic regression more effectively and efficiently. Genetic programming (GP) and weighted K-nearest neighbours (KNN) are the foundations of this approach. To anticipate the missing values of partial features, it builds GP-based models utilizing additional features that are accessible. Weighted KNN is used to choose the instances that are utilized to create these types of models.

Cubillos et al. [18] provide a bi-objective technique based on the k-nearest neighbours (biokNN) approach for missing value imputation. The approach improves imputation accuracy and helps lessen multilevel model bias, mainly when there is a high intraclass correlation and a high missing rate. An approach for missing value imputation utilizing clustering and linear regression was proposed by Karmitsa et al. [19]. They only employ data points to estimate missing values compared to incomplete data points. Imputation techniques based on clustering use a similar idea. However, the linear regression technique is used inside each cluster to predict missing values accurately. Thomas and Rajabi [20] present a descriptive overview of machine learning (ML) applications for imputing missing values connected to experimental settings, including datatypes, missingness processes, platforms, missing ratios, and dataset characteristics.

2.2 Outlier Detection

Du et al. [21] suggested two new outlier identification algorithms for non-numerical datasets. It described an Outlier Detection Tree (ODT) technique based on entropy. To divide the data set into two classes—a normal class and an aberrant class—a classification tree is built using ODT. If-then rules in the tree are used in each data object is classified as either a normal or an outlier. An advanced outlier identification technique is developed for excellent detection precision and minimal time complexity.

The Z-order curve is used by Ma et al. [22] to determine the kNN in their weighted kNN query approach. The process first computes the weight of each attribute using information entropy, then encodes high-dimensional data into Z-value using the Z-order curve. Every object's weighted kNN is searched based on its Z-value. Concurrently, a new technique for detecting outliers is introduced, which utilizes the weighted kNN of each object and its average and minimum distances from it.

A mean-shift outlier detector is suggested by Yang et al. [23]. This detector modifies the data and eliminates

the bias produced by the outliers using the mean-shift technique. By substituting the mean of each item for its k-nearest neighbours, the mean-shift technique effectively removes the impact of outliers before clustering without requiring knowledge of the outliers. Additionally, it uses the distance shifted to identify outliers.

Li et al. [24] proposed a weighted outlier mining technique to find outliers in multidimensional categorical datasets. It comprises two unique modules: 1) feature grouping using feature correlation measurements and 2) outlier mining using score assignments to individual feature group objects. Li et al. [25] proposed an attribute-weighted outlier detection approach for multidimensional and massive mixed data. An attribute-weighting method is provided for diverse data, and the algorithm uses mutual information to determine feature correlations.

2.3 Imbalanced Data Classification

The imbalance problem is addressed by a unique online learning technique from incomplete and imbalanced data streams [26]. Its main idea is in two parts: 1) it finds the most valuable attributes in incomplete feature spaces by applying the empirical risk minimization principle, and 2) by converting F-measure optimization into a weighted surrogate loss minimization, it creates an adaptive cost approach for managing imbalanced class distributions in real-time.

Edward et al. [27] created a rebalancing system for multi-classifying unbalanced healthcare information using SMOTE and Cluster-based techniques. Dynamic Ensemble Selection for multi-class for enhanced multi-classification performance, Shapley Additive explanations and Recursive Feature Elimination feature selection method, and sampling method to correct the unbalanced class distribution.

A two-stage cost-sensitive classification method is suggested to tackle class imbalance in non-stationary data streams. Sun et al. [28] propose a two-stage, cost-sensitive approach for classifying data streams using cost information in both the feature selection and classification phases. Additionally, this approach incorporates a window adaptation and drift detection mechanism that ensures an ensemble can quickly adjust to idea drift.

By utilizing the new weighted vote parameters for the weak classifiers, Wang et al. [29] suggest a way to enhance the AdaBoost algorithm. The global error rate and the positive class's classification accuracy rate influence the weighted vote parameters. One further consideration in developing the algorithms is the data's unbalanced index. Tang et al. [30] developed a hybrid framework that combines feature selection and ensemble-based learning. This process consists of three stages: gathering information about the unbalanced data in the first stage, lowering the high dimensionality of the data in the second stage, and feature selection in the third stage, which entails choosing the most pertinent features.

III. PROPOSED METHODOLOGY

This section explains the proposed pre-processing techniques to improve the data quality. Figure 3.1 shows the proposed architecture diagram.

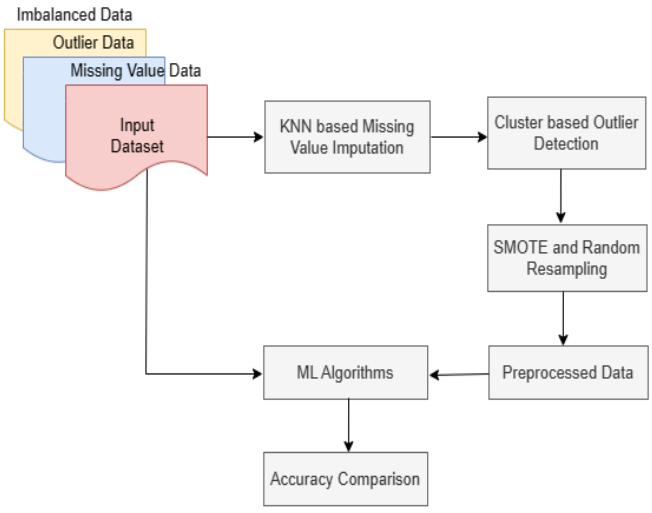

Figure 3.1 Proposed Work Architecture

3.1 KNN-based Missing Value Imputation

Generally, dataset D contains a set of data with class label represented by, D = {(X_1, Y_1), (X_2, Y_2),…., (X_n, Y_n)}, where n is the total number of instances, X_i ={f_1,f_2,f_3,…,f_m}, m indicates the total number of attributes and Y_i ={c_1,c_2,c_3,…, cc}, where c is the number of unique class of the dataset. The question mark (?) represents the missing values. Three main mechanisms can be used to define missing values [31]: missing not at random (MNAR) denotes that a variable's missingness is dependent on itself; missing completely at random (MCAR) happens if the missingness presented in medical data follows an entirely random pattern; and missing at random (MAR) signifies a variable's missingness is related to another variable. In datasets, missing values are quite undesirable. Designing a suitable approach and procedures to deal with the missing values is necessary. Algorithm 1 explains the proposed KNN and distance-based missing value imputation method.

Algorithm-1 KNN-based Missing Value Imputation

Input: Dataset D = {(X_1,Y_1), (X_2,Y_2),…., (X_n,Y_n)} contains some missing values at random
Output: Dataset DMVI (missing value imputed)
Step01: For i = 1 to m
Step02: mc_i = compute the missing percentage of the i^{th} column using Eq. (1)
Step03: If $mc_i > 0.40$ then
Step04: Remove the i^{th} column
Step05: End For
Step06: Split the D into D_{wom} (without missing Instances) and D_{wim} (with missing Instances)
Step07: Group D_{wom} and D_{wim} based on the class label
Step08: for each Instance in D_{wim}
Step09: K_{inst} = find the K nearest neighbour instance from D_{wom}
Step10: If the missing attribute f_j is categorical, then
Step11: Assign high frequency
Step12: If the missing attribute f_j is numerical, then
Step13: Assign the mean value
Step14: End For

Initially, compute the missing rate for each column in the dataset using,

$$mc_i = \frac{\sum_{j=1}^{m} v_{ji}}{m}, \quad \begin{cases} v_{ji} = 1, & if\ D_{j,i} = '?' \\ v_{ji} = 0, & Otherwise \end{cases}$$

If the missing rate of the particular column is more significant than 0.40, remove the specific column. Divide the dataset into D_{wom} (without missing Instances) and D_{wim} (with missing Instances) and group the D_{wom} and D_{wim} based on the class label. The proposed algorithm uses the KNN technique to consider the distance in the dataset's space between the sample vectors while imputeing missing data. It averages the values of the k nearest samples with this feature observed for each missing feature in numerical data. The outcome for categorical data is the class that occurs the most frequently among the k nearest neighbours. The Euclidean is the metric used to calculate the distance between two Instances, p and q:

$$dist(X_p, X_q) = \sum_{j=1}^{m} (X_{pj} - X_{qj})^2 \quad (2)$$

Where m is the number of features, X_{pj} is the j^{th} feature of the Instance of X_p, and X_{qj} is the j^{th} feature of the Instance of X_q.

Consider the m_{ij} the missing value of the j^{th} feature of i^{th} Instance. The missing value can be imputed using,

$$m_{ij} = \begin{cases} \frac{1}{k}\sum_{t=1}^{k} m_{tj}, & if\ m_{ij}\ is\ numerical\ feature \\ \max_{k}(m_{tj}), & if\ m_{ij}\ is\ categorical\ feature \end{cases} \quad (3)$$

3.2 Cluster-based Outlier Detection

This section explains the cluster-based outlier detection algorithm. Typically, clustering techniques analyze the similarity between data points using distance metrics (Euclidean distance) and combine similar data points into clusters that exhibit similar behaviour. The data points are regarded as outliers if

they are either not part of any clusters or are part of clusters much smaller than others.

Algorithm 2 explains the cluster-based outlier detection

Algorithm-2 Cluster-based outlier detection

Input: Dataset D = {(X_1,Y_1), (X_2,Y_2),…., (X_n,Y_n)}
Output: Outlier Dataset OD
Step01: som = Apply SOMCluster(D,c)
Step02: somEval = Get ClassesToCluster(som)
Step03: For each Instance inst in somEval
Step04: If inst.classLable != inst.clusterLable
Step05: initOut.add(inst)
Step06: EndFor
Step07: For each Cluster C_i in Cls
Step08: Compute radius of cluster R(C_i) using Eq. (4)
Step09: Compute local cluster outlier factor LCOF(C_i) using Eq.(5)
Step10: EndFor
Step11: For each Instances out1 in initOut
Step12: Compute Instance outlier factor IOF(out1) using Eq. (6)
Step13: If IOF(out1) > LCOF(C_i) (out1 ϵ C_i) then
Step14: OD.add (out1) [set out1is outlier]
Step15: EndFor

Initially, the SOM (Self Organization Map) algorithm is used to cluster the dataset—the traditional unsupervised learning neural network model known as SOM groups comparable input data into clusters. The SOM reduces the dimensionality of information by combining the projection and clustering methods. The result of the clustering algorithm is evaluated using the classes to clusters evaluation method. The initial outliers are extracted based on the classes to clusters evaluation.

The average distance between member objects and the centroid is the cluster radius. The average distance squared between a cluster's instances and centroid. It can be computed using the following Eq.

$$R(C_i) = \sqrt{\frac{\sum_{j=1}^{cn}\left(inst_j - cent_i\right)^2}{cn}}$$

Where cn is the number of instances in cluster C_i. $inst_j$ is the j^{th} Instance of cluster C_i, and $cent_i$ indicates the centroid of the cluster C_i

This paper computes each cluster's local cluster outlier factor using the following Eq.

$$LCOF(C_i) = \frac{\max_{j=1:instances(C_i)} dist(inst_j, Cent_i)}{R(C_i)} \quad (5)$$

The instance outlier factor can be computed using,

$$IOF(inst) = \frac{\max_{j=1:instances(C_i)} dist(inst_j, Cent_i)}{dist(inst, Cent_i)} \quad (6)$$

If the IOF(inst) < $LCOF(C_i)$ Where inst ϵ C_i, then the inst is an outlier

3.3 SMOTE and Random Resample imbalance data handling

Reducing the number of instances in more classes or increasing the number in minority classes are two examples of data-level methods utilized in the training data to make the class distribution more balanced. These methods can alter the dataset structure to the greatest extent possible to balance the imbalanced class. There are three resampling methods: under-sampling, over-sampling and hybrid approach. When using under-sampling techniques, instances from the majority class are eliminated until each class has almost the same amount of instances. In over-sampling techniques, class boundaries are reinforced while achieving an equal sample distribution by creating new samples based on samples from the minority class. The hybrid sampling approach combines under- and over-sampling. This paper uses SMOTE and a random resampling approach to rebalance the imbalanced dataset. The SMOTE is used

for over-sampling, and random resampling is used for under-sampling.

IV. EXPERIMENTAL RESULTS

The performance evaluation of the research activity is explained in this section. The proposed data pre-processing model was simulated and evaluated using Java and the recommended experimental configuration. The proposed pre-processing approach was assessed using real-time medical datasets collected from the UCI and Kaggle datasets. Table 4.1 shows the dataset description.

Table 4.1 Dataset Description

Dataset	# Instances	# Features	# Missing Values	# Outliers	Class Distribution
Cirrhosis	412	17	973	-	c1 = 21, c2= 92, c3 = 155, c4 = 144
Diabetes	768	8	-	268	c0 = 500, c1 = 268
Dermatology	366	34	8	-	c1= 112, c2 = 61, c3 = 72, c4 = 49, c5 = 52, c6 = 20
Ecoli	336	7	27	9	cp = 143, im = 77, imS = 2, imL = 2, imU = 35, om = 20, omL = 5, pp = 52
HepatitisC	615	12	31	-	c1 = 540, c2 = 24, c3 = 21, c4 = 30
Storke	5110	10	201	-	c0 = 3550, c1 = 249

The following metrics are used to analyze the performance of the proposed pre-processing approach.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (10)$$

Where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

The classification algorithms Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF) and Support Vector Machine (SVM). Table 4.2 shows the accuracy of the original dataset (without pre-processing).

Table 4.2 Accuracy without Pre-processing Methods

Algorithms	Cirrhosis	Dermatology	Diabetes	Ecoli	HepatitisC	Stroke
KNN	41.99	95.35	70.18	81.54	90.89	91.85
NB	43.20	**97.54**	76.30	**84.52**	91.05	88.61
RF	48.54	96.44	74.34	84.22	**93.17**	94.18
SVM	**51.21**	97.26	**77.34**	82.73	90.89	**95.12**

Figure 4.1 shows the accuracy comparison for different medical datasets without pre-processing. From the results, Cirrhosis, Diabets, and Storke datasets achieved 51.21%, 77.34%, and 95.12%

accuracy for the SVM algorithm, and Dermatology and Ecoli datasets attained 97.54% and 84.52% accuracy for the NB algorithm. The Hepatitis C dataset achieves 93.17% accuracy for the RF algorithm.

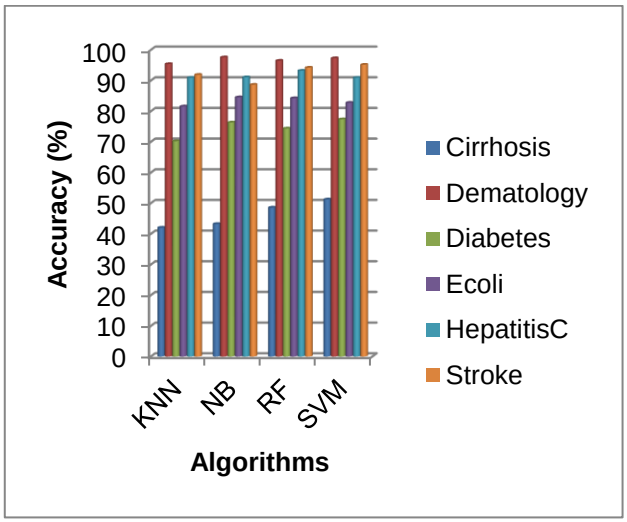

Figure 4.1 Accuracy Comparison without Pre-processing

Table 4.3 and Figure 4.2 show the accuracy comparison after missing value imputation.

Table 4.3 Accuracy after missing value imputation

Algorithms	Cirrhosis	Dermatology	Diabetes	Ecoli	Hepatitis C	Stroke
KNN	47.81	95.35	70.18	83.33	90.56	92.29
NB	51.69	**97.54**	76.30	**86.01**	91.54	89.33
RF	**60.68**	96.72	74.34	80.35	**93.98**	94.75
SVM	53.15	96.99	**77.34**	82.14	91.05	**95.12**

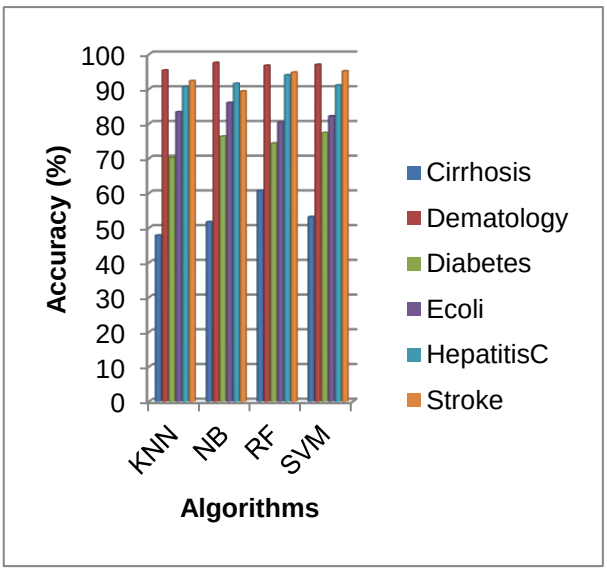

Figure 4.2 Accuracy Comparison after missing value imputation

From the results, Cirrhosis, the Hepatitis C dataset attains 60.68% and 93.98% accuracy for the RF algorithm, and the Dermatology and Ecoli datasets gain 97.54% and 86.01% accuracy for the NB algorithm. The diabetes and Stroke dataset achieves 77.34% and 95.12% accuracy for the SVM algorithm. Table 4.4 and Figure 4.3 show the accuracy comparison after outlier removal.

Table 4.4 Accuracy after outlier removal

Algorithms	Cirrhosis	Dermatology	Diabetes	Ecoli	HepatitisC	Stroke
KNN	51.63	**99.24**	76.38	96.92	93.99	94.68
NB	49.81	98.47	81.45	97.94	**95.06**	91.48
RF	**57.09**	98.85	79.87	**98.97**	94.63	95.78
SVM	**57.09**	**99.24**	**82.40**	**98.97**	93.13	**96.05**

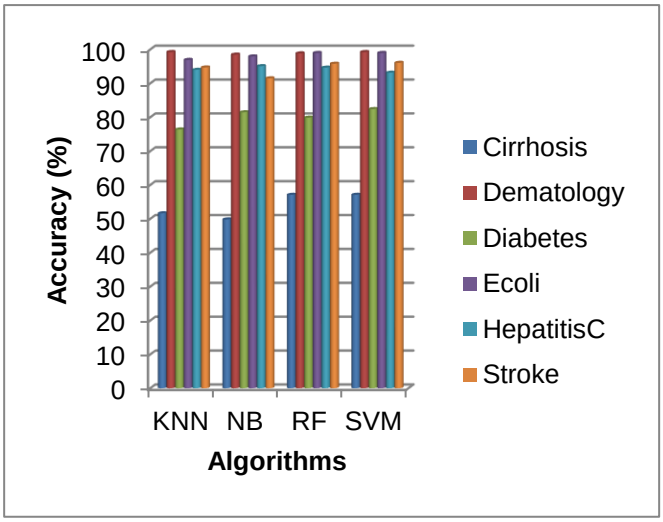

Figure 2.3 Accuracy Comparison after Outlier Removal

From the results, the SVM algorithm provides high accuracy 57.09%, 99.24%, 82.40%, 98.97%, and 96.05% for Cirrhosis, Dermatology, Diabetes, Ecoli, and Stroke datasets. The HepatitisC dataset attains 95.06% for the NB algorithm.

Table 4.5 and Figure 4.4 show the accuracy comparison after over-sampling. The datasets Dermatology, Diabetes, Ecoli achieve 97.87%, 82.43% and 98.98% for the SVM algorithm. Cirrhosis and Stroke attain 60.99% and 95.65 for RF, and Hepatitis C has 95.12% for the NB algorithm.

Table 4.5 Accuracy after Over-Sampling

Algorithms	Cirrhosis	Dermatology	Diabetes	Ecoli	HepatitisC	Stroke
KNN	56.38	96.54	80	96.44	94.06	93.48
NB	52.12	95.74	80.27	97.97	**95.12**	90.77
RF	**60.99**	97.07	81.21	98.47	94.91	**95.65**
SVM	56.38	**97.87**	**82.43**	**98.98**	91.73	93.88

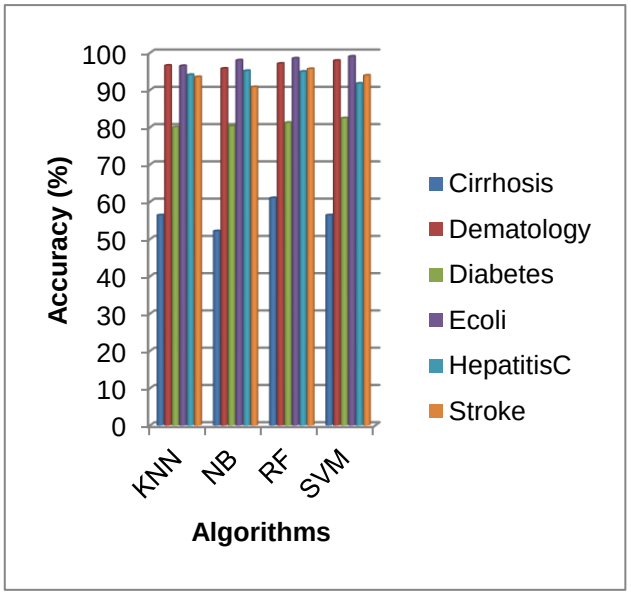

Figure 4.4 Accuracy Comparison after Over-Sampling

Table 4.6 and Figure 4.5 show the accuracy comparison after under-sampling. Cirrhosis, Dermatology, Hepatitis C and Stroke achieve 71.53%, 97.81%, 97.42% and 97.26% for the RF algorithm. The dataset Diabetes attains 86.98% for SVM, and Ecoli has 97.93% for the NB algorithm.

Table 4.6 Accuracy after under-sampling

Algorithms	Cirrhosis	Dermatology	Diabetes	Ecoli	HepatitisC	Stroke
KNN	60.58	97.26	80.95	91.75	95.70	96.62
NB	57.66	97.26	83.17	**97.93**	93.99	91.82
RF	**71.53**	**97.81**	85.07	96.90	**97.42**	**97.26**
SVM	56.93	96.72	**86.98**	94.84	92.27	95.91

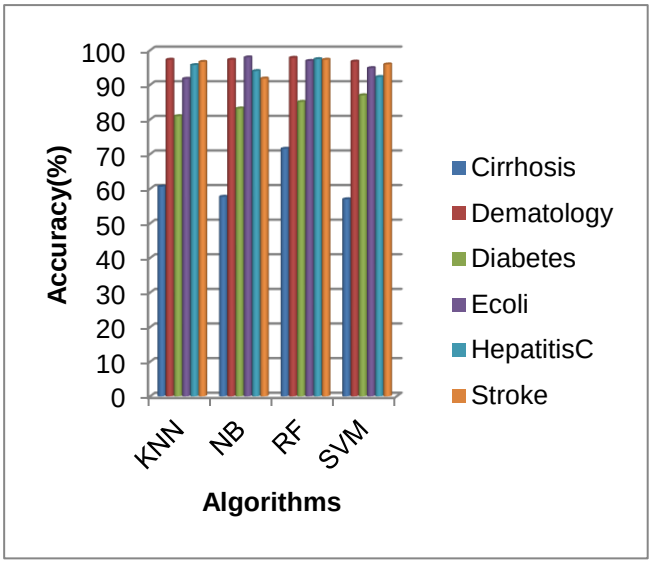

Figure 4.5 Accuracy Comparison after Under-Sampling

V. CONCLUSIONS

Pre-processing data is an essential step in the data mining process. However, if there is a lot of noisy and erroneous data or irrelevant and redundant information, data analysis during the learning phase becomes more difficult. This study suggests effective pre-processing strategies to enhance the effectiveness of the learning process and the quality of the data. Imputation of missing values, outlier detection, and treatment of unbalanced data are all included in the pre-processing procedure. The study outcomes demonstrate that the suggested pre-processing technique improved overall performance.

III. REFERENCES

[1]. Almuhaideb, S., & Menai, M. E. B. (2016). Impact of pre-processing on medical data classification. Frontiers of Computer Science, 10, 1082-1102.

[2]. Idri, A., Benhar, H., Fernández-Alemán, J. L., & Kadi, I. (2018). A systematic map of medical data pre-processing in knowledge discovery. Computer methods and programs in biomedicine, 162, 69-85.

[3]. Jena, M., & Dehuri, S. (2022). An Integrated Novel Framework for Coping Missing Values Imputation and Classification. IEEE Access, 10, 69373-69387.

[4]. Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524.

[5]. Lin, W. C., Tsai, C. F., & Zhong, J. R. (2022). Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems, 239, 108079.

[6]. Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends, 1(2), 56-70.

[7]. Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. Big Data and Cognitive Computing, 5(1).

[8]. Orooji, A., & Kermani, F. (2021). Machine learning based methods for handling imbalanced data in hepatitis diagnosis. Frontiers in Health Informatics, 10(1), 57.

[9]. Psychogyios, K., Ilias, L., Ntanos, C., & Askounis, D. (2023). Missing value imputation methods for electronic health records. IEEE Access, 11, 21562-21574.

[10]. Nijman, S. W. J., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J. L., Bots, M. L., (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. Journal of clinical epidemiology, 142, 218-229.

[11]. Le, T. D., Beuran, R., & Tan, Y. (2018). Comparison of the most influential missing data

imputation algorithms for healthcare. In 2018 10th international conference on knowledge and systems engineering (KSE) (pp. 247-251). IEEE.

[12]. Samara, M. A., Bennis, I., Abouaissa, A., & Lorenz, P. (2022). A survey of outlier detection techniques in IoT: review and classification. Journal of Sensor and Actuator Networks, 11(1), 4.

[13]. Christy, A., Gandhi, G. M., & Vaithyasubramanian, S. (2015). Cluster based outlier detection algorithm for healthcare data. Procedia Computer Science, 50, 209-215.

[14]. Palli, A. S., Jaafar, J., Hashmani, M. A., Gomes, H. M., & Gilal, A. R. (2022). A hybrid sampling approach for imbalanced binary and multi-class data using clustering analysis. IEEE Access, 10, 118639-118653.

[15]. Ofek, N., Rokach, L., Stern, R., & Shabtai, A. (2017). Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. Neurocomputing, 243, 88-102.

[16]. Zhang, X., Yan, C., Gao, C., Malin, B. A., & Chen, Y. (2020). Predicting missing values in medical data via XGBoost regression. Journal of healthcare informatics research, 4, 383-394.

[17]. Al-Helali, B., Chen, Q., Xue, B., & Zhang, M. (2021). A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data. Soft Computing, 25, 5993-6012.

[18]. Cubillos, M., Wøhlk, S., & Wulff, J. N. (2022). A bi-objective k-nearest-neighbors-based imputation method for multilevel data. Expert Systems with Applications, 204, 117298.

[19]. Karmitsa, N., Taheri, S., Bagirov, A., & Mäkinen, P. (2020). Missing value imputation via clusterwise linear regression. IEEE Transactions on Knowledge and Data Engineering, 34(4), 1889-1901.

[20]. Thomas, T., & Rajabi, E. (2021). A systematic review of machine learning-based missing value imputation techniques. Data Technologies and Applications, 55(4), 558-585.

[21]. Du, H., Ye, Q., Sun, Z., Liu, C., & Xu, W. (2020). FAST-ODT: A lightweight outlier detection scheme for categorical data sets. IEEE Transactions on Network Science and Engineering, 8(1), 13-24.

[22]. Ma, Y., & Zhao, X. (2021). POD: a parallel outlier detection algorithm using weighted KNN. IEEE Access, 9, 81765-81777.

[23]. Yang, J., Rahardja, S., & Fränti, P. (2021). Mean-shift outlier detection and filtering. Pattern Recognition, 115, 107874.

[24]. Li, J., Zhang, J., Pang, N., & Qin, X. (2020). Weighted outlier detection of high-dimensional categorical data using feature grouping. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(11), 4295-4308.

Cite this article as :

Prof. Arati K Kale, Dr. Dev Ras Pandey, "Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 11 Issue 1, pp. 299-309, January-February 2024. Available at doi : https://doi.org/10.32628/IJSRST52411130

Journal URL : https://ijsrst.com/IJSRST52411130