International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

An Automated Progressive Data Cleaning Framework for Lung Cancer Medical Data using Machine Learning

B. Samirana Acharya*[1], K. Ramasubramanian[2]

Submitted: 08/05/2023 **Revised**: 15/07/2023 **Accepted**: 06/08/2023

Abstract: With the immense growth in the field of computational algorithms and data management, the demand for automating the medical analysis and diagnosis is also increasing. The foundational demand from the medical analysis is rapid analysis with least error or almost with zero errors. The manual process is subjected to the higher human interventions and with higher scope of errors. Henceforth, dealing with analysis of life treating diseases such as lung cancer must be automated. The challenge with the computer driven automated processes is the quality of the data decides the accuracy of the final outcomes or information. Henceforth, the data cleaning or as called literally data pre-processing is one of the major focused areas of concern for building automated frameworks for disease detections. Many Researchers have dedicatedly worked towards achieving the best pre-processing framework. Nonetheless, these research attempts are criticised for various reasons such not designed for medical information pre-processing as various parameters like precision, “missing value” and dimension of the data plays a major role. Few of parallel research outcomes have demonstrated higher focus on the medical information pre-processing while building the framework. However, these methods demonstrate higher complexity and hard to adapt due to strong dependency on the “dataset”. Henceforth, the paper proposes a novel framework for medical data pre-processing with few benchmarking proposed algorithms with adaptive and threshold driven method for “outlier” detection and imputation, domain specific “missing value” detection and imputation, and finally mete information specific noise reduction. The outcome of the proposed framework demonstrates nearly 50% improvement with the benchmarked algorithms attached with the proposed framework due to this adaptation.

Keywords*: Medical Data Pre-processing, Lung Canter data pre-processing, outlier detection, noise reduction, missing value imputations*

1. Introduction

The automation for disease detection is one of the prime research domains for various research attempts. The availability of multiple algorithms for data pre-processing and processing for data made the researchers to produce higher level of effectiveness from the data management frameworks. Nonetheless, the medical data is critical due to various factors as highlighted in the work by N. Nasrullah et al. [1] and demands special treatments during pre-processing. With the inclusion of machine learning algorithms in the computing algorithms domain, the processing of the medical data has seen tremendous benefits. The work by I. Ali et al. [2] have demonstrated the applicability of machine learning methods for lung cancer data pre-processing and processing with significant outcomes. The pre-processing phase for machine learning must be designed with the focus on the mechanisms and algorithms to be used for processing the data for actual outcomes such as classification or clustering or prediction. The work by W. Zuo et al. [3] showcases the adaptation of phase specific pre-processing methods for disease detection on lung cancer data.

The automation for disease detection using machine learning algorithms have been the highlight for the past few years. Nonetheless, the automation on the pre-processing of the medical data is the demand of the recent research trends as suggested in the work by N. Gupta et al. [4]. The bottlenecks for building such frameworks are as followings:

- The medical data are extremely sensitive during the pre-processing operations such as “outlier” detections as many data items can be identified as “outlier” by the framework and removal of such data points can be disastrous in terms of information loss.
- Also, the medical data demands higher precision and few of the higher precision data points can be detected as noisy data. Nonetheless, reduction of the precision from such data points can reduce the accuracy of the further stages of medical data processing.
- Finally, the “missing value” imputation is highly critical for medical data as the traditional methods such as imputation with mean value cannot be applied. The imputation must consider various other parameter values for the same patient.

[1] *Research Scholar, Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India*
ORCID ID : 0000-0002-8852-4567
[2] *Associate professor Department of Computer science and Engineering Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India*
ORCID ID : 0000-0002-7163-1251
** Corresponding Author Email: samiranaacharya@gmail.com*

Henceforth, this work proposes a novel framework for automating the data pre-processing specific to the medical data.

The rest of the paper is organized such as in the Section – II, III and IV the foundational strategies "outlier", "missing value" and noise detection for generic data are discussed. Based on the foundational understanding, the critical analysis of the parallel research outcomes is carried out in the Section – V. The persistent challenges in research attempts are discussed in the Section – VI. The proposed solutions mathematical models and the proposed algorithms are furnished in Section – VII and VIII respectively. The obtained results are discussed in the Section – IX and these results are again compared with the benchmarked parallel research outcomes in the Section – X and the work presents the research conclusion in the Section – XI.

2. "Outlier" Detection Fundamentals

After setting the context for the research in the previous section of this work, in this section, the foundational method for "outlier" detection is analyzed.

Assuming that, the complete "dataset" is DS[] and each and every attribute can be denoted as A_X. Hence for a total of n number of attributes the relation can be formulated as,

$$DS[] = < A_1, A_2, A_3, ...A_n > \quad \text{(Eq. 1)}$$

Further for each of attribute the data domain is consisting of individual data points as D_X. Hence for m number of data points in each data domain, the relation can be formulated as,

$$A_X = < D_1, D_2, D_3, ...D_m > \quad \text{(Eq. 2)}$$

As per the foundational strategy, the mean of the data domain must be calculated to be the driving threshold for identification of the "outlier" value. Hence the calculation of the threshold, as $A_X.D_{th}$, must be realized first as,

$$A_X.D_{th} = \frac{\sum_{i=1}^{m} D_i}{\Phi(A_X[])} \quad \text{(Eq. 3)}$$

Where, Φ denotes the count of the data items in the domain.

Finally, the detection of the "outlier" data item is realized using the threshold value as,

$$Outlier[] \leftarrow \begin{cases} If, D_X[i] > A_X.D_{th} \\ Else, Not\ Outlier \end{cases} \quad \text{(Eq. 4)}$$

The challenges in this foundation method are discussed in the further sections of this work.

After the realization of the foundational method for "outlier" detection, in the next section, the "missing value" detection fundamentals are discussed.

3. "Missing Value" Detection Fundamentals

After the realization of the "outlier" detection foundation in the previous section of this work, in this section yet another challenge for data pre-processing is considered and realized as "missing value" identification.

Non-response may result in missing data: no information is supplied for one or more components, or for the whole unit, as a result of not responding. There are certain questions that are more likely than others to elicit a nonresponse, such as ones that pertain to private matters such as money. Attrition is a sort of "missing value" that may occur in longitudinal studies—for example, when researching development when a measurement is repeated after a particular amount of time and is defined as the absence of data from a study. Missing Data point happens when a participant drops out of the test before it is completed, resulting in one or more measures being missing.

Continuing from the Eq. 2, the foundational method demonstrates very simple solution to identify the "missing value" as for any data points in the "dataset", if the value is null, denoted as null, then that specified data point must be considered as "missing value".

This can be formulated as,

$$MV[] \leftarrow \begin{cases} If, D_X[i] == null \\ Else, Normal\ Data\ Item \end{cases}$$

(Eq. 5)

The challenges in this foundation method are discussed in the further sections of this work.

After the realization of the foundational method for "outlier" detection, in the next section, the data noise detection fundamentals are discussed.

4. Data Noise Detection Fundamentals

After the realization of the "missing value" detection foundation in the previous section of this work, in this section yet another challenge for data pre-processing is considered and realized as data with noise identification.

Data with noise as well as data that has a poor Signal-to-Noise Ratio, is referred to as noisy data. Improper processes (or procedures that are not adequately documented) for removing noise from data might result in a misleading impression of accuracy or incorrect findings.

Noisy data are data that contains a significant quantity of extra worthless information, which is referred to as noise. Among the things that are included is data corruption, and the phrase is often used as a synonym for faulty data. It also includes any information that a user system is unable to comprehend and interpret effectively. Unstructured text, for example, is incompatible with many systems. If noisy data is not handled appropriately, it may have a negative impact on the outcomes of any data study and can distort conclusions. Statistical analysis is occasionally used to separate the signal from the noise in a set of data.

The foundational method clearly states that, the metadata information, if meta{X} function extracts the meta data information from the data domain, must be valid for each and every data item and failing to match, the data items must be identified as noisy data point.

The meta function can be formulated as,

$$meta\{D_X[]\} \rightarrow T_X \qquad \text{(Eq. 6)}$$

Here, T_X is the extracted meta data type from the domain of the data. The detection of the noisy data as stated by the foundational method, can be formulated as,

$$Noise[] \leftarrow \begin{Bmatrix} If, D_X[i] \neq T_X \\ Else, No\ Noise \end{Bmatrix} \qquad \text{(Eq. 7)}$$

The challenges in this foundation method are discussed in the further sections of this work.

After the realizing the foundational method for noise detection in the numeric information, in the next section, the parallel research outcomes as recent improvements over the foundational methods are critically discussed.

5. Parallel Research Outcomes

After realizing the foundational methods over the previous three sections of this work, in this section, the recent improvements over these foundational methods are critically analyzed.

The process of detecting any disease, specifically the lung cancers, is to separate the normal patent information from the "dataset" and the rest of the available information can be categorized as diseased data as showcased in the work by Y. Chen et al. [5] using the neural networks for clustering. The adaptation of the neural networks for the clustering process is highly appreciated by many researchers for identification of the influencing factors to determine the cancer possibilities with the automatic corrections of neural network node weights can reduce the time complexity to a greater extend. Nonetheless, the actual available data is in the image format collected from magnet sources such as MR images. Hence, these data demand a higher level of pre-processing to be applied to any machine learning algorithms or neural networks as clearly demonstrated in the work by G. Wei et al. [6].

The recent work demonstrated by J. Gong et al. [7] have highlighted yet another possibility of applying higher order machine learning algorithms such as 3D models on the MR images. It is natural to realize that the time complexity of such model will be significantly higher compared to the standard machine learning algorithms, hence a form of reduction to the "dataset" are highly expected, which again can be achieved by pre-processing on the "dataset".

Many of the medical researchers believe that the lung cancer as a disease is influenced by genetics of the patient. Hence, while designing the algorithms for lung cancer detection various research attempts have shown the inclusion of genome sequences in the processing of the patient data. One such example can be seen in the work by J. J. Chabon et al. [8]. This increases the complexity of the model to a great extend and demands further reduction of the characteristics of the "dataset" with substantial cleaning in terms of "missing value" and "outlier". In the other hand, such strategies demand higher memory requirements as the considering data have doubled in size. This complexity aspect is highlighted in the work by A. Mobiny et al. [9].

Fundamentally, the standard process for detecting the lung cancers can be performed using the standard parameters extracted from the MR images or from the textual "dataset" extracted from the patient information as demonstrated by M. A. Heuvelmans et al. [10]. Some of the parallel research outcomes have also demonstrated the use of CT scans as source of the data. Nevertheless, this form of data is highly prune to the human errors and demands higher pre-processing in terms of noise reductions as clearly indicated in the work by I. W. Harsono et al. [11].

After the initial phases of data accumulation, the further part of such processes is to build the knowledge bases from the data. During the knowledge base building process, the "dataset" are expected to be complete with all the "missing value" and "outlier" replaced as strongly recommended by Y. Xie et al. [12] and I. Ali et al. [13]. This proposed work, demonstrates a nearly accurate regression-based imputation of the "missing value" and "outlier". Also, during the phase of building the knowledge base, the data pre-processing modules in every algorithm has to implement huge data. It is natural to realize that the complexity of the actual algorithm can be reduced with cleaner data, however, the pre-processing algorithms add a significant amount of time complexity to the existing algorithm. To overcome the additional time complexity, many of the parallel research attempts have demonstrated the use neural network driven algorithms to reduce the time complexity one such example can be seen in the works by I. Ali et al. [14], A. Naik et al. [15] , C.-J. Lin et al. [16], and R. Dey et al. [17].

In the other hand, the hardware driven solutions are also popular for detecting CT scan-based lung cancers. The work by M. Al-Shabi et al. [18] demonstrated a solution towards this approach. Nonetheless, this method is highly criticised for higher dependencies on the CT scan devices and for the huge noises captured during this process. The solution to this problem can be the adaption of machine learning algorithms and the detection of the lung cancer can be achieved using only the scan data from CT machines as suggested by R. V. M. D. Nobrega et al. [19].

Speaking about the accuracy of these parallel research outcomes, the accuracies obtained from these methods are arguable as majority of the algorithms have demonstrated tightly coupled pre-processing algorithms very specific to the "dataset". The work by Y. Qin et al. [20] have formulated a critical analysis about these works and highlighted that the pre-processing phases must be independent to the "dataset" and the clustering or detection algorithms. The same was supported in the works by D. Ardila et al. [21] and P. Monkam et al. [22].

Fundamentally, apart from the actual clustering or classification methods to detect the lung cancer, the processing of the raw information set relies on the image processing fundaments as suggested by B. Fielding et al. [23] , C. Zhao et al. [24] and B. X. Chen et al. [25]. Nonetheless, this proposed work assumes that the data for processing are already extracted from image data and available in the textual formats.

Henceforth, the conclusive statements from the literature survey are:

- The generic pre-processing blocks the demand of the current research for further improvements over the detection accuracy.
- The pre-processing blocks are to be attached separately to the framework in order to bring higher flexibility towards the adaptation of various clustering and classification algorithms.
- The model complexity of the pre-processing phases in the framework must be less.

Henceforth, in the next section of this work, the identified problems are furnished using mathematical models.

6. Problem Formulation

After realizing the persisting challenges in the current research outcomes, this section of the work is dedicated to formulation of the same challenges using mathematical models. These mathematical models further will help us in realizing the proposed solutions in further sections of this work.

Continuing from the Eq. 3 and assuming the data points are D_X and D_Y with the following conditions as,

$$D_X < D_Y \qquad \text{(Eq. 8)}$$

And,

$$D_X < A_X .D_{th} \qquad \text{(Eq. 9)}$$

And,

$$D_Y < A_X .D_{th} \qquad \text{(Eq. 10)}$$

Assuming that, the different between the mean value and the datapoints are D1 and D2 respectively, then this can be formulated as,

$$D1 = \left| D_X - A_X .D_{th} \right| \quad \text{(Eq. 11)}$$

And,

$$D2 = \left| D_Y - A_X .D_{th} \right| \quad \text{(Eq. 12)}$$

Clearly, D1-> 0 and thus D_X is identified as not "outlier" and D_Y is identified as "outlier" as per the existing proposals.

However, in case of the following relation, the identification of D_Y as "outlier" is again questionable.

$$\left| D1 \right| - \left| D2 \right| \approx 0 \qquad \text{(Eq. 13)}$$

In the existing systems, this is the primary challenge as the detection of the "outlier" are not relative and these methods are only applicable to the numeric data points.

Further, during the processing of the "missing value" or the "outlier", the complexity T(n) of the existing system can be calculated in continuation to the Eq. 2 as,

$$T(n) = n * m \qquad \text{(Eq. 14)}$$

Or,

$$T(n) = n * n, n \approx m \quad \text{(Eq. 15)}$$

As,

$$T(n) = O(n^2) \qquad \text{(Eq. 16)}$$

Which is significantly higher considering the higher values of n.

Finally, the imputation method as per the parallel research outcomes are using the mean values for "outlier" and "missing value" and sometimes in case of noisy data as well.

Henceforth, this proposed work elaborates the solutions in order to solve these problems, which are again furnished in the next section of this work.

7. Proposed Solutions: Mathematical Models

Firstly, the "outlier" detection and imputation methods are furnished.

The proposed method, identifies the domain mean in a progressive strategy as initially on the first iteration it identifies the mean as,

$$A_X .D_{th} = \frac{\sum_{i=0}^{1} D_i}{\Phi(\left| A_X [] \right|_0^1)} \quad \text{(Eq. 17)}$$

And, for the kth iteration, the domain mean can be calculated as,

$$A_X.D_{th} = \frac{\sum_{i=0}^{m-k} D_i}{\Phi(\left|A_X[]\right|_0^k)} \quad \text{(Eq. 18)}$$

And, as stated in Eq. 4, the identification of the "outlier" can be realized with the help of domain mean.

The proposed method also recommends the imputation of the "outlier" value as following,

$$D_X[i] = \beta_0 + \sum_{j=0}^{n-X} \beta_j.D_j \qquad \text{(Eq. 19)}$$

The above stated regression-based method for calculating the imputation value considers the other parameters values, i.e. from 0 to n-X. This strategy not only resolves the "outlier" issues, rather also imputes the value with most meaningful manner possible.

Secondly, for the "missing value" detection and imputation methods, this work analyses the length of the domain for successive domains. Assuming that, two successive domains such as D_X and D_Y, the length of the domains are L1 and L2, respectively, can be calculated as following:

$$L1 = \Phi(D_X[]) \qquad \text{(Eq. 20)}$$

And,

$$L2 = \Phi(D_Y[]) \qquad \text{(Eq. 21)}$$

The strategy defined in this work to identify any number of "missing value" is if the length of the domains is different, then either of the domain contains the "missing value" and the domain with the lower length contains the "missing value" as identified. This relation can be formulated as following,

$$if \left| \begin{array}{l} L1 > L2,\ D_Y[] \to MV \\ L1 < L2,\ D_X[] \to MV \end{array} \right. \qquad \text{(Eq. 22)}$$

Further, the domain, identified with the "missing value" (MV), must be processed further. Assuming that, the domain D_X, contains the "missing value". Thus, the domain must be partitioned in initially two parts as (0 to r) and (r to m) in terms of number of observations and the Eq. 22 is applied successively for identification of the "missing value".

Once the "missing value" is identified, the imputation method will be applied as stated in Eq. 19.

Thirdly and finally, the noisy data identification and imputation methods are formulated. Continuing from Eq. 7, the detected noisy data can be identified as, $D_X[i]$. Further, the imputation method realizes the type, T_X, and precision, P_X, for the data domain D_X as following,

$$Meta\left|D_X[]\right| \Rightarrow T_X : T_X\left\{D_X[i]\right\} = P_X \qquad \text{(Eq. 23)}$$

Further for any data point, the precision type does not match with the extracted precision, then the imputation method is applied as,

$$if\ T_X \neq P_X \left|D_X[i] \leftarrow T_X\left\{D_X[i]\right\}\right. \qquad \text{(Eq. 24)}$$

The imputation method in this work for noisy data is formulated as, the precision of the actual domain data type will be applied to the noisy data to reduce the noise without losing the precision.

Henceforth, with the detailed elaboration on the proposed methods using mathematical models, the next section of this work elaborates the proposed algorithms based on these formulations.

8. Proposed Algorithms and Framework

After the detailed analysis of the existing problems in the research and the proposed solutions using the mathematical models in the previous section of this work, in this section, the proposed algorithms and the framework are established.

Firstly, the regression-based imputation value calculation algorithm is furnished.

Algorithm - I: Imputation Value Calculation using Regression Coefficients (**IVC-RC**)

Input:

- The "dataset" record as DR[X][]
- Replaceable Value as RV

Output:

- Imputation value as V

Process:

```
Step - 1. Accept the "dataset" record as DR[X][]
Step - 2. For each data point in DR[X][i]
        a. Calculate the regression coefficient as R[i]
           using Eq. 19
        b. Replace RV using the Eq. 19
Step - 3. Return V
```

Secondly, the "outlier" detection and imputation algorithm is furnished.

Algorithm - II: Progressive "outlier" Detection and Imputation using Regression Method (**POD-IRM**)

Input:

- The raw "dataset" as DS[][]

Output:

- “outlier” reduced / removed “dataset” as DS1[][]

Process:

```
Step - 1. Load the “dataset” as DS[][]
Step - 2. For each attribute domain in DS[][] as DS[i][]
        a. Calculate the progressive mean as Mean[i] using Eq. 18
        b. If DS[i][j] > Mean[i]
            i. Then, Call IVC-RC algorithm with DS[i][j]
            ii. Replace DS[i][j] in DS1[][] with V from IVC-RC algorithm
Step - 3. Test the Accuracy of Clustering using standard KMeans & BIRCH & DBSCAN
Step - 4. Return DS1[][]
```

An “outlier” is a data point that stands out from the rest. It may be related to measurement variability or experimental error, which is occasionally eliminated from the data set. An “outlier” may skew statistical results.

“outlier” may occur in any distribution by chance, but they frequently suggest measurement error or a heavy-tailed population. This indicates that the distribution has considerable skewness and therefore tools or intuitions that presume a normal distribution should be used with caution. A mixing of two distributions, which may represent two separate sub-populations, or may reflect "accurate trial" against "measurement mistake", is a common source of “outlier”.

Further, the “missing value” detection and imputation algorithm is furnished.

Algorithm - III: Progressive “missing value” Detection and Imputation using Regression Method (**PMV-IRM**)

Input:

- “outlier” reduced / removed “dataset” as DS1[][]

Output:

- “missing value” reduced / removed “dataset” as DS2[][]

Process:

```
Step - 1. Load the “dataset” as DS1[][]
Step - 2. For each attribute domain in DS1[][] as DS1[i][]
        a. Calculate the length of the domain as Len[i] and Len[i+1] using Eq. 21
        b. Identify the presence of “missing value” using Eq. 22
        c. If Len[i] contains “missing value”
            i. Then repeat from Step - 2 with DS1[i][]
        d. Else, repeat from Step - 2 with DS1[i+1][]
        e. Identify the “missing value” as MV with DS1[i][j]
        f. Call IVC-RC algorithm with DS1[i][j]
        g. Replace DS1[i][j] in DS2[][] with V from IVC-RC algorithm
        h. Test the Accuracy of Clustering using standard KMeans & BIRCH & DBSCAN
Step - 3. Return DS2[][]
```

Finally, the noisy data detection and imputation algorithm is furnished here.

Algorithm - IV: Noisy Data Detection and Imputation using Regression Method (**NDD-IRM**)

Input:

- “missing value” reduced / removed “dataset” as DS2[][]

Output:

- Noisy data reduced / removed “dataset” as DS3[][]

Process:

```
Step - 1. Load the “dataset” as DS2[][]
Step - 2. For each attribute domain in DS2[][] as DS2[i][]
        a. Extract the meta data information from using Eq. 23
        b. Identify the presence of meta information mismatch using Eq. 24
        c. Replace the format using Eq. 24
Step - 3. Return DS3[][]
```

Further, the proposed framework is furnished here [Fig - 1].

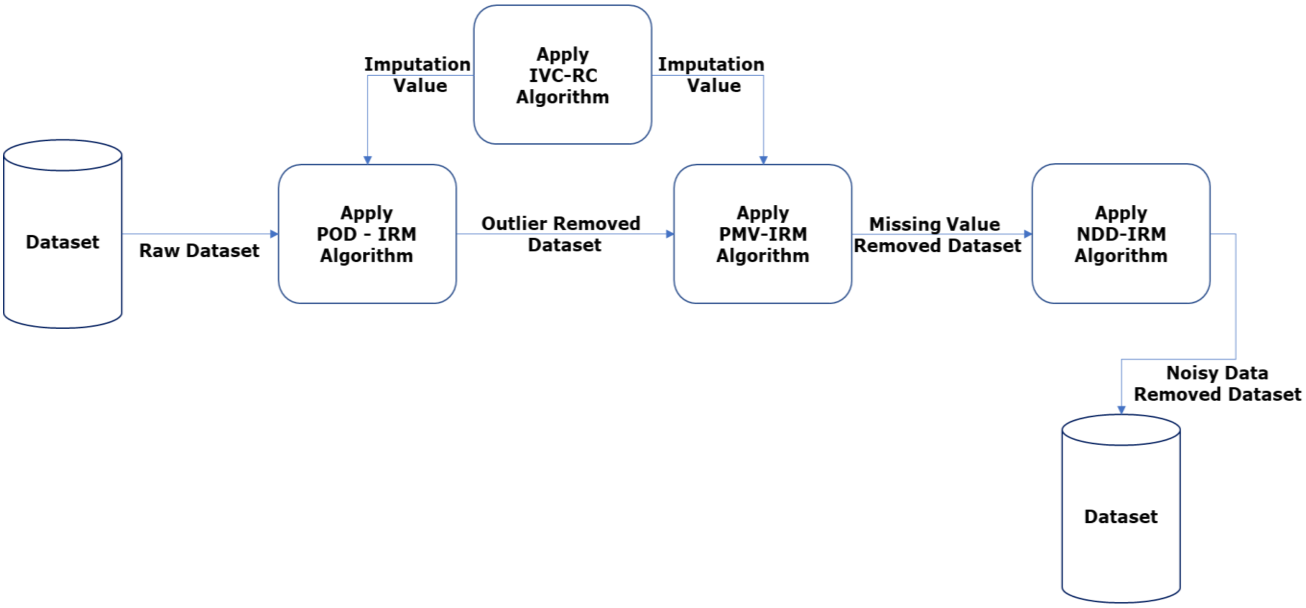

Fig. 1. Proposed Data Pre-Processing Framework

Further, the next section discusses the obtained results from the proposed framework.

9. Results and Discussions

After the detailed analysis of the proposed methods using the mathematical model and further analyzing the algorithms, in this section the obtained results are reported.

Firstly, the "dataset" initial analysis is furnished here [Table – 1]

Table 1: Initial "Dataset" Analysis

Characteristics	***Value***
Total Number of Attributes	25
Total Number of Records	1000
Total Number of "outlier"	11606
Total Number of "missing value"	429

It is observable that, the "dataset" is highly distributed with good number of "outlier" and "missing value".

Secondly, the distribution of the "dataset" is analyzed again with max-min comparisons [Table – 2] for few attributes.

Table 2: "dataset" Distribution Analysis

Attribute Name	***count***	***mean***	***std***	***min***	***25%***	***50%***	***75%***	***max***
Patient Id	1000	500.5	288.819436	1	250.75	500.5	750.25	1000
Age	1000	37.174	12.005493	14	27.75	36	45	73
Gender	1000	1.402	0.490547	1	1	1	2	2
Air Pollution	1000	3.84	2.0304	1	2	3	6	8
Alcohol use	1000	4.563	2.620477	1	2	5	7	8
Dust Allergy	1000	5.165	1.980833	1	4	6	7	8
OccuPational Hazards	1000	4.84	2.107805	1	3	5	7	8
Genetic Risk	1000	4.58	2.126999	1	2	5	7	7
chronic Lung Disease	1000	4.38	1.848518	1	3	4	6	7
Balanced Diet	1000	3.856	2.244616	1	2	3	5	9
Weight Loss	720	4.576389	2.214442	1	3	5	7	8

Attribute Name	***count***	***mean***	***std***	***min***	***25%***	***50%***	***75%***	***max***
Shortness of Breath	1000	4.24	2.285087	1	2	4	6	9
Wheezing	851	4.26322	1.819838	2	2	4	6	8
Swallowing Difficulty	1000	3.746	2.270383	1	2	4	5	8
Clubbing of Finger nails	1000	3.923	2.388048	1	2	4	5	9
Frequent Cold	1000	3.536	1.832502	1	2	3	5	7
Dry Cough	1000	3.853	2.039007	1	2	4	6	7
Snoring	1000	2.926	1.474686	1	2	3	4	7
Patient Id	1000	500.5	288.819436	1	250.75	500.5	750.25	1000
Age	1000	37.174	12.005493	14	27.75	36	45	73

The distribution is again analyzed graphically here [Fig – 2].

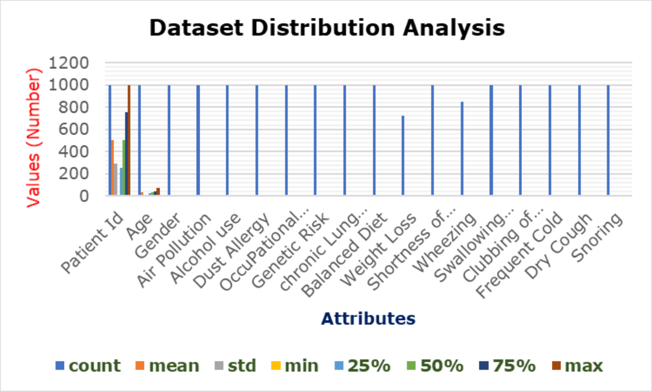

Fig. 2.“dataset” Distribution Analysis

Here from the graphical analysis, it is notable that, the “missing value” and the “outlier” are available in few of the attribute domains.

Henceforth, in the next phase, the “missing value” detection algorithm is applied [Table – 3].

Table 3: “missing value” Analysis

Characteristics	***Number of “missing value”***	***% of “missing value”***	***Number of “missing value” Imputed***
Patient Id	0	0%	0
Age	0	0%	0
Gender	0	0%	0
Air Pollution	0	0%	0
Alcohol use	0	0%	0
Dust Allergy	0	0%	0
OccuPational Hazards	0	0%	0
Genetic Risk	0	0%	0
chronic Lung Disease	0	0%	0
Balanced Diet	0	0%	0
Obesity	0	0%	0
Smoking	0	0%	0
Passive Smoker	0	0%	0
Chest Pain	0	0%	0
Coughing of Blood	0	0%	0
Fatigue	0	0%	0
Weight Loss	280	28%	280
Shortness of Breath	0	0%	0
Wheezing	149	15%	149
Swallowing Difficulty	0	0%	0
Clubbing of Finger Nails	0	0%	0
Frequent Cold	0	0%	0
Dry Cough	0	0%	0
Snoring	0	0%	0
Level	0	0%	0

Further, the outcomes are analyzed graphically [Fig – 3].

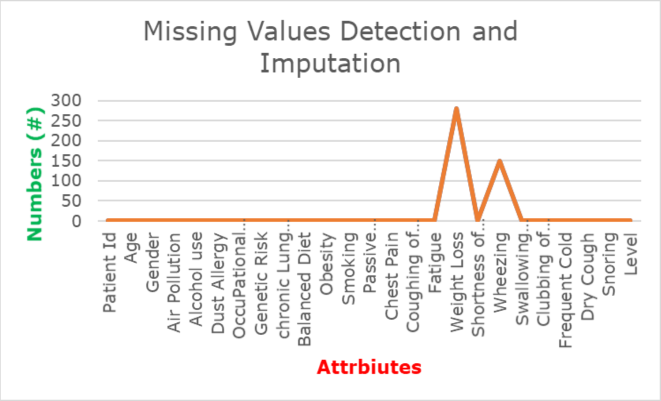

Fig. 3.“missing value” Detection and Imputation Analysis

Henceforth, the proposed algorithm has detected and imputed 100% “missing value” from the “dataset”.

Further, the “outlier” detection and imputation algorithm is applied on the “dataset” and the obtained results are furnished here [Table – 4].

Table 3: “outlier” Detection and Imputation Analysis

Characteristics	***Number of “outlier”***	***% of “outlier”***	***Number of “outlier” Imputed***
Patient Id	500	50%	500
Age	431	43%	431
Gender	402	40%	402
Air Pollution	485	49%	485
Alcohol use	525	53%	525
Dust Allergy	525	53%	525
OccuPational Hazards	555	56%	555
Genetic Risk	535	54%	535
chronic Lung Disease	495	50%	495
Balanced Diet	495	50%	495
Obesity	406	41%	406
Smoking	425	43%	425
Passive Smoker	355	36%	355
Chest Pain	395	40%	395
Coughing of Blood	465	47%	465
Fatigue	467	47%	467
Weight Loss	389	39%	389
Shortness of Breath	447	45%	447
Wheezing	388	39%	388
Swallowing Difficulty	529	53%	529
Clubbing of Finger Nails	529	53%	529
Frequent Cold	439	44%	439
Dry Cough	529	53%	529
Snoring	530	53%	530
Level	365	37%	365

Further, the results are visualized graphically here [Fig – 4].

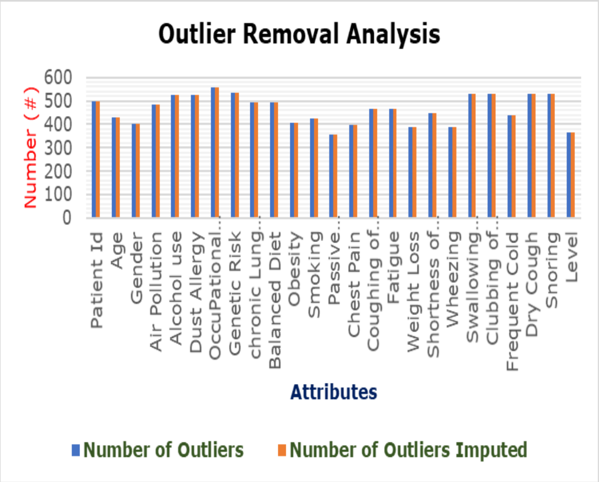

Fig. 4.“outlier” Values Detection and Imputation Analysis

Finally, the cleaned “dataset” was passed to various clustering algorithms, before and after pre-processing. The results are furnished here [Table – 5].

Table 4: Clustering Accuracy Analysis

Characteristics	***Accuracy On Raw “dataset” (%)***	***Accuracy After “missing value” Imputations (%)***	***Accuracy After “outlier” Imputations (%)***
K-Means	33.40	49.90	89.90
DB-Scan	24.00	41.85	88.31
BIRCH	20.01	38.04	72.76

Characteristics	***Accuracy On Raw “dataset” (%)***	***Accuracy After “missing value” Imputations (%)***	***Accuracy After “outlier” Imputations (%)***
Agglomerative Clustering	31.35	36.83	87.04

The final analysis results are also visualized graphically here [Fig – 5].

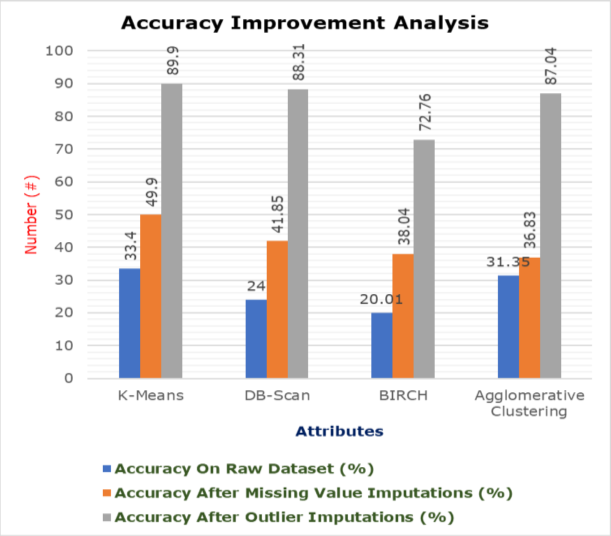

Fig. 5. Accuracy Improvement Analysis

Henceforth, it is natural to realize that the proposed framework has improved the accuracy of the standard methods almost by 50%.

The tested “dataset” is free from all noisy data, hence, the last algorithm is not applied to the “dataset”.

Further, the proposed framework is compared with the other parallel research outcomes in the next section.

10. Comparative Analysis

After analyzing the obtained results in the previous section of this work, in this section, the proposed framework outcomes are compared with other parallel research outcomes [Table – 6].

Table 5: lustering Accuracy Analysis

Author, Year	***Pre-Processing Method***	***Model Complexity***	***Accuracy (%)***
A. Mobiny et al., 2021 [10]	Neural Network	O(n^2)	76.03
M. A. Heuvelmans et al., 2021 [11]	Deep Learning	O(n^2)	77.08
A. Naik et. al., 2021 [15]	Max – Min Method	O(n^2)	77.12
C. Zhao et al., 2020 [24]	Max – Min Method	O(n^2)	78.23
Proposed Framework	Machine Learning	O(n)	84.50

Again, it is natural to observe that, the proposed framework has outperformed the other parallel research outcomes with least complexity during pre-processing of the “dataset”.

Finally, in the next section of the work, the research conclusion is presented.

11. Conclusion

The need for automating medical analysis and diagnosis is rising in tandem with the expansion of computer methods and data management. Rapid analysis with few or no mistakes is a basic need for medical analysis. Manual processes need more human interaction and are more prone to mistakes. The examination of life-threatening illnesses like lung cancer must now be computerised. The issue with computer-driven automated procedures is that the data quality determines the end results or information. Henceforth, data cleansing or pre-processing is one of the key focal areas for constructing automated disease detection systems. In recent years, several concurrent research efforts have been made to find the optimal pre-processing framework. Nonetheless, these research efforts are criticised for not being built for medical data pre-processing, where factors like accuracy, “missing value”, and data dimension are crucial. The pre-processing of medical data is emphasised in certain parallel research outputs. Due to the reliance on the “dataset”, these approaches are more sophisticated and difficult to adjust. Now this study presents a unique framework for medical data pre-processing with adaptive and threshold driven “outlier” identification and imputation, domain specific “missing value” detection and imputation, and noise reduction. This adaption has resulted in a roughly 50% improvement in the benchmarked algorithms tied to the proposed framework.

References

[1] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai and H. Hu, "Automated lung nodule detection and classification using deep learning combined with

multiple strategies", Sensors, vol. 19, no. 17, pp. 3722, Aug. 2019.

[2] I. Ali, G. R. Hart, G. Gunabushanam, Y. Liang, W. Muhammad, B. Nartowt, et al., "Lung nodule detection via deep reinforcement learning", Frontiers Oncol., vol. 8, pp. 108, Apr. 2018.

[3] W. Zuo, F. Zhou, Z. Li and L. Wang, "Multi-resolution CNN and knowledge transfer for candidate classification in lung nodule detection", IEEE Access, vol. 7, pp. 32510-32521, 2019.

[4] N. Gupta, D. Gupta, A. Khanna, P. P. R. Filho and V. H. C. de Albuquerque, "Evolutionary algorithms for automatic lung disease detection", Measurement, vol. 140, pp. 590-608, Jul. 2019.

[5] Y. Chen, Y. Wang, F. Hu and D. Wang, "A lung dense deep convolution neural network for robust lung parenchyma segmentation", IEEE Access, vol. 8, pp. 93527-93547, 2020.

[6] A. M. Anter and A. E. Hassenian, "CT liver tumor segmentation hybrid approach using neutrosophic sets fast fuzzy C-means and adaptive watershed algorithm", Artif. Intell. Med., vol. 97, pp. 105-117, Jun. 2019.

[7] G. Wei, H. Cao, H. Ma, S. Qi, W. Qian and Z. Ma, "Content-based image retrieval for lung nodule classification using texture features and learned distance metric", J. Med. Syst., vol. 42, no. 1, pp. 13, Jan. 2018.

[8] J. Gong, J.-Y. Liu, L.-J. Wang, X.-W. Sun, B. Zheng and S.-D. Nie, "Automatic detection of pulmonary nodules in CT images by incorporating 3D tensor filtering with local image feature analysis", Phys. Medica, vol. 46, pp. 124-133, Feb. 2018.

[9] J. J. Chabon, E. G. Hamilton, D. M. Kurtz, M. S. Esfahani, E. J. Moding, H. Stehr, et al., "Integrating genomic features for non-invasive early lung cancer detection", Nature, vol. 580, pp. 245-251, Apr. 2020.

[10] A. Mobiny, P. Yuan, P. A. Cicalese, S. K. Moulik, N. Garg, C. C. Wu, et al., "Memory-augmented capsule network for adaptable lung nodule classification", IEEE Trans. Med. Imag., Jan. 2021.

[11] M. A. Heuvelmans, P. M. A. van Ooijen, S. Ather, C. F. Silva, D. Han, C. P. Heussel, et al., "Lung cancer prediction by deep learning to identify benign lung nodules", Lung Cancer, vol. 154, pp. 1-4, Apr. 2021.

[12] I. W. Harsono, S. Liawatimena and T. W. Cenggoro, "Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning", J. King Saud Univ.-Comput. Inf. Sci., vol. 1319, pp. 1-8, Apr. 2020.

[13] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham, et al., "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT", IEEE Trans. Med. Imag., vol. 38, no. 4, pp. 991-1004, Apr. 2019.

[14] I. Ali, M. Muzammil, I. U. Haq, A. A. Khaliq and S. Abdullah, "Efficient lung nodule classification using transferable texture convolutional neural network", IEEE Access, vol. 8, pp. 175859-175870, 2020.

[15] A. Naik and D. R. Edla, "Lung nodule classification on computed tomography images using deep learning", Wireless Pers. Commun., vol. 116, pp. 655-690, Jan. 2021.

[16] C.-J. Lin and Y.-C. Li, "Lung nodule classification using Taguchi-based convolutional neural networks for computer tomography images", Electronics, vol. 9, no. 7, pp. 1066, Jun. 2020.

[17] R. Dey, Z. Lu and Y. Hong, "Diagnostic classification of lung nodules using 3D neural networks", Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), pp. 774-778, Apr. 2018.

[18] M. Al-Shabi, H. K. Lee and M. Tan, "Gated-dilated networks for lung nodule classification in CT scans", IEEE Access, vol. 7, pp. 178827-178838, 2019.

[19] R. V. M. D. Nobrega, S. A. Peixoto, S. P. P. D. Silva and P. P. R. Filho, "Lung nodule classification via deep transfer learning in CT lung images", Proc. IEEE 31st Int. Symp. Comput. Based Med. Syst. (CBMS), pp. 244-249, Jun. 2018.

[20] Y. Qin, H. Zheng, Y. M. Zhu and J. Yang, "Simultaneous accurate detection of pulmonary nodules and false positive reduction using 3D CNNs", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 1005-1009, Apr. 2018.

[21] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, et al., "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography", Nature Med., vol. 25, no. 6, pp. 954-961, Jun. 2019.

[22] P. Monkam, S. Qi, H. Ma, W. Gao, Y. Yao and W. Qian, "Detection and classification of pulmonary nodules using convolutional neural networks: A survey", IEEE Access, vol. 7, pp. 78075-78091, 2019.

[23] B. Fielding and L. Zhang, "Evolving deep DenseBlock architecture ensembles for image classification", Electronics, vol. 9, no. 11, pp. 1880, Nov. 2020.

[24] C. Zhao, T. F. Wang and B. Y. Lei, "Medical image fusion method based on dense block and deep convolutional generative adversarial network", Neural Comput. Appl., vol. 11600, pp. 1-16, Oct. 2020.

[25] B. X. Chen, T. J. Liu, K. H. Liu, H. H. Liu and S. C. Pei, "Image super-resolution using complex dense block on generative adversarial networks", Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 2866-2870, Sep. 2019.

[26] Sherje, N. P., Agrawal, S. A., Umbarkar, A. M., Kharche, P. P., & Dhabliya, D. (2021). Machinability study and optimization of CNC drilling process parameters for HSLA steel with coated and uncoated drill bit. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1070

[27] Moore, B., Clark, R., Muñoz, S., Rodríguez, D., & López, L. Automated Grading Systems in Engineering Education: A Machine Learning Approach. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/125

[28] Gandhi, L. ., Rishi, R. ., & Sharma, S. . (2023). An Efficient and Robust Tuple Timestamp Hybrid Historical Relational Data Model. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 01–10. https://doi.org/10.17762/ijritcc.v11i3.6193