CHAPTER 4

ADDITIVE LOG RATIO AND ZERO MEAN FEATURE NORMALIZED ENCODING BASED DATA PREPROCESSING FOR DISEASE IDENTIFICATION

4.1 INTRODUCTION

Disease prediction refers to a process detecting the possibility of a patient's disease after examining the integrations of the patient's symptoms. Early detection of mortality, considering the symptoms that appear in patients with disease, is important. In recent years, artificial intelligence techniques have been widely applied in the medical field to enhance the efficiency of disease diagnosis and classification in their early stages. Among those techniques, machine learning (ML) techniques stand out, which are a set of statistical models that assist the machine learn from past data. Despite this, it is often challenging to analyse patient data for diagnosis in the early stages due to factors such as large data volumes, missing values, and data noise. Despite its shortcomings, machine learning models and their capabilities have enabled the processing of such data.

Also, it is noticeable that data features may be incomplete and huge. The range of some data features varies from small to big. The type of data features combines both categorical and numerical; this will affect the accuracy of ML techniques in diagnosing and classifying diseases in their early stages. Applying data preprocessing techniques to manage the data feature increases the accuracy of ML models in the early detection of the disease. Data preprocessing is the primary process of data mining. It converts data into a model-friendly format, enabling efficient data processing.

Preprocessing is the process of cleaning and organizing the text for an accurate classification process. Data preprocessing is crucial in ensuring the data is disease-free. Medical datasets are often unbalanced and contain missing values; therefore, the performance of classifiers cannot be efficiently achieved without addressing these issues. Data preprocessing helps to eliminate noise and missing values, providing efficient results in disease classification.

If there is a huge quantity of improper and superfluous information present or noisy and incompatible data, then knowledge discovery during the training time may be further complicated. Data preparation and filtering concepts take a substantial sum of processing time. Examples of methods used in data preprocessing include instance selection, normalisation, and data transformation, among others. The primary goal of data preprocessing is to enhance the quality of the data and make it more suitable for subsequent processing. Many existing data preprocessing methods have been studied to improve the prediction performance.

The proposed preprocessing techniques are compared with existing methods, such as the Convolutional Neural Networks-Gated Recurrent Unit (CNN-GRU) based hybrid deep learning model, to predict the number of deaths and analyse the spread in popular countries. Another method, the Variant Of Concerned Deep Learning (VOC-DL) prediction framework, was developed to forecast the pandemic. However, the space complexity and time complexity during preprocessing were not minimized by the above-mentioned methods. Therefore, a novel and efficient preprocessing technique is needed for better disease prediction.

The Additive Log Ratio Transformed One-Hot Encoding (ALRTOHE) Technique is proposed to perform data preprocessing in a minimal amount of time with maximum accuracy. First, the amount of data from the given dataset is taken as input. With this input data, preprocessing is performed using the ALRTOHE technique in two phases: additive log-ratio transformation and one-hot encoding. In the first phase, an additive log-ratio transformation is employed to normalise the data within a specific range. Second, data decoding is performed to convert the numerical data from binary coding. Here, the data decoding is accomplished for modifying numerical data within binary coding. The ALRTOHE technique utilises one-hot encoding to transform numerical categorical variables into binary vectors. From that, the binary representation of data is attained in data preprocessing in less time. The results of the proposed ALRTOHE technique are compared with those of conventional methods in terms of preprocessing accuracy, space complexity, and preprocessing time.

The Zero Mean Feature Normalised Encoding (ZMFNE) technique is proposed for preprocessing the dataset with higher accuracy. Based on the data normalisation process and

encoding process, the ZMFNE technique is employed to perform data preprocessing. In this phase, unnecessary and inconsistent data are eliminated to minimise complexity during processing. The input data samples and features are taken from the dataset. Based on the input data and features, the sample input matrix is generated in rows and columns. Then, zero-mean feature scaling is employed to normalise the input data. Furthermore, the data transformation process is carried out using one-hot encoding. The encoder modifies the categorical features into a numeric array. The encoder takes the integer array-like strings as input. Here, features are encoded and generate a binary column for each category. From this, the data normalisation and transformation process is carried out, aiding in achieving the preprocessed data output. Experiments of the proposed ZMFNE technique are conducted on factors such as preprocessing accuracy, pre-processing time and space complexity. The performance analysis of the proposed ZMFNE technique improves accuracy while preprocessing in less time and with minimal complexity.

4.2 PROPOSED ADDITIVE LOG RATIO TRANSFORMED ONE HOT ENCODING-BASED DATA PREPROCESSING

Data preprocessing is the primary step applied to clean and change the raw data set into a structured format. Typically, the dataset contains missing and duplicate values in an unusable format. Machine learning models do not directly consider the dataset to undergo the prediction process. The missing data in the dataset affects the accuracy of the task. Additionally, the process of working with such a raw dataset introduces time complexity. This emphasises the need to preprocess the input dataset into a digestible format. With this intent, the Additive Log Ratio Transformed One-Hot Encoding (ALRTOHE) technique is proposed to preprocess the input data with minimal time and space complexity for disease prediction.

The data preprocessing using the ALRTOHE technique is performed to handle insufficient and inconsistent data in the dataset. The normalisation process is carried out to normalise various scales of categorical attributes by mapping them to an integer value. When diverse attributes are there- attributes having values on various scales- it leads to the performance of classification operations. Hence, all the attributes are initially normalized to obtain the same

scale values. Additionally, data preprocessing is used to eliminate noise and decrease memory space. The block diagram of ALRTOHE-based data preprocessing is shown in Figure 4.1.

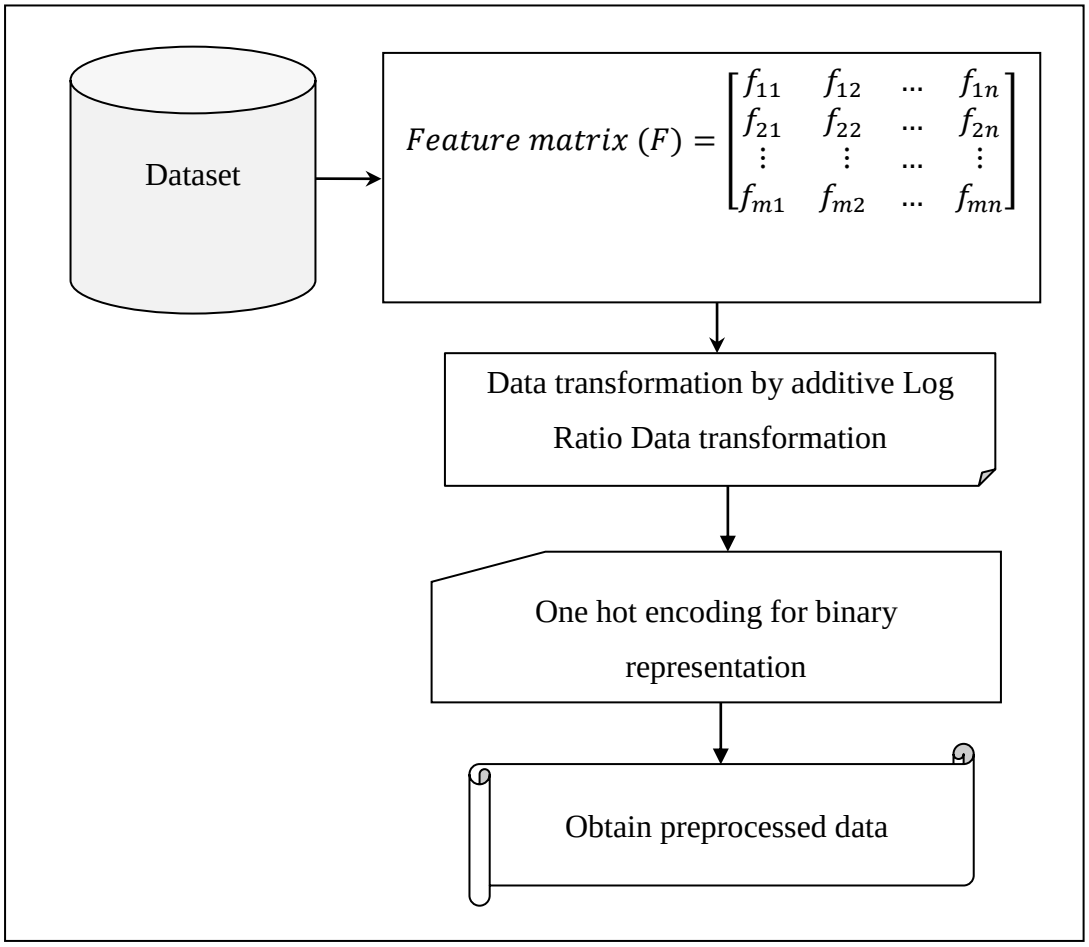


Figure 4.1 Block diagram of Additive Log Ratio Transformed One Hot Encoding based Data Preprocessing

Figure 4.1 depicts the preprocessing process based on the ALRTOHE technique. The proposed ALRTOHE technique gathers the patient information from the dataset. By considering this dataset, data preprocessing is performed to identify the patients infected by the disease accurately. An elaborate explanation of the proposed ALRTOHE technique is as follows.

Let us consider a dataset 'D' that includes of features '$F = f_1, f_2, .., f_m$'. Initially, data normalization is performed using additive log-ratio transformation and the one-hot encoding technique. Additive log-ratio transformation is a method that rescales the attribute values into 0 and 1 range.

Let us assume the feature matrix with 'n' rows and 'm' columns in the dataset. The additive log ratio with standard deviation is employed to normalize the data, as shown.

$$LR_a = \left(\frac{\log\left|f_{vi-m_f}\right|}{D_s}\right) \tag{4.1}$$

Where,

$$D_s = \sqrt{\frac{\left(f_{vi-m_f}\right)^2}{n}} \tag{4.2}$$

Where 'LR_a' points out an output of Additive Log Ratio results, 'f_{vi}' points out a feature value, 'm_f' represents a mean of the particular feature value, 'D_s' represents a standard deviation, 'n' refers the number of samples. The output of LR_a gives attribute values in the ranges 0 and 1.

Later, data decoding is obtained to transfer the numerical data within the binary coding. The proposed ALRTOHE technique utilizes One Hot Encoding which aids in modify numerical categorical variables into binary vectors. Before implementing the normalized data into an algorithm, make sure that all the numerical attribute values are encoded. Assume the normalization of features in the array of matrix 'NA' in equation (4.3).

$$NF = \begin{bmatrix} Nf_{11} & Nf_{12} & \dots & Nf_{1n} \\ Nf_{21} & Nf_{22} & \dots & Nf_{2n} \\ \vdots & \vdots & \dots & \vdots \\ Nf_{m1} & Nf_{m2} & \dots & Nf_{mn} \end{bmatrix} \tag{4.3}$$

Where 'NF' indicates a normalized feature matrix, and '$Nf_{11}, Nf_{12} \dots, Nf_{1n}$' refers the normalized numerical value of the features. Followed by, the input numerical feature value is then encoded to obtain the binary value using equation (4.4).

$$OHE \xleftarrow{NBR} \begin{bmatrix} Nf_{11} & Nf_{12} & \dots & Nf_{1n} \\ Nf_{21} & Nf_{22} & \dots & Nf_{2n} \\ \vdots & \vdots & \dots & \vdots \\ Nf_{m1} & Nf_{m2} & \dots & Nf_{mn} \end{bmatrix} \tag{4.4}$$

0

Where 'NBR' indicates the numeric to binary representation with Equation (4.5).

$$OHE \xrightarrow{returns} \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \dots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} \quad (4.5)$$

Where 'OHE' gives the binary representation 'b' as an output. It helps to minimize the time and space of ALRTOHE technique. The algorithmic process of additive log-ratio transformed one hot encoding-based data preprocessing is described as given below.

// Algorithm 4.1: Additive Log Ratio transformed One Hot Encoding based Data Preprocessing
Input: Dataset D, features $F = f_1 f_2, .., f_m$'
Output: Obtain the pre-processed data
Begin **step 1**: Collect the number of features from the dataset 'D' **step 2**: **For** each feature in the dataset 'f_i' **step 3:** Construct the feature matrix 'A' **step 4:** Measure the log ratio with mean and standard deviation 'LR_a' **step 5:** Normalize the data in the ranges from 0 to 1 **step 6:** **End for** **step 7:** **Apply** OHE to a normalized data **step 8:** **Return** (binary representation of data '0' or '1') **step 9: Obtain the preprocessed dataset** **End**

Algorithm 4.1 illustrates the data preprocessing process aimed at reducing the time and space complexity metrics. Initially, a collection of features and raw data are obtained as input from a given data repository. Following this, an additive log-ratio transformation is applied to get the normalised data. Subsequently, OHE is employed for transforming the quantitative data into

binary values. Therefore, preprocessed data is obtained in a binary representation to reduce the space complexity.

4.3 PROPOSED ZERO MEAN FEATURE NORMALIZED ENCODING TECHNIQUE

Another technique, called Zero Mean Feature Normalised Encoding (ZMFNE), has been developed to preprocess the input data with higher accuracy in less time. The primary benefits of the ZMFNE Technique are to clean data and make it suitable for classification, thereby enhancing the accuracy and effectiveness of the model. The proposed ZMFNE technique performs data pre-processing for normalisation and encoding. The process of the ZMFNE Technique for data processing is depicted in Figure 4.2.

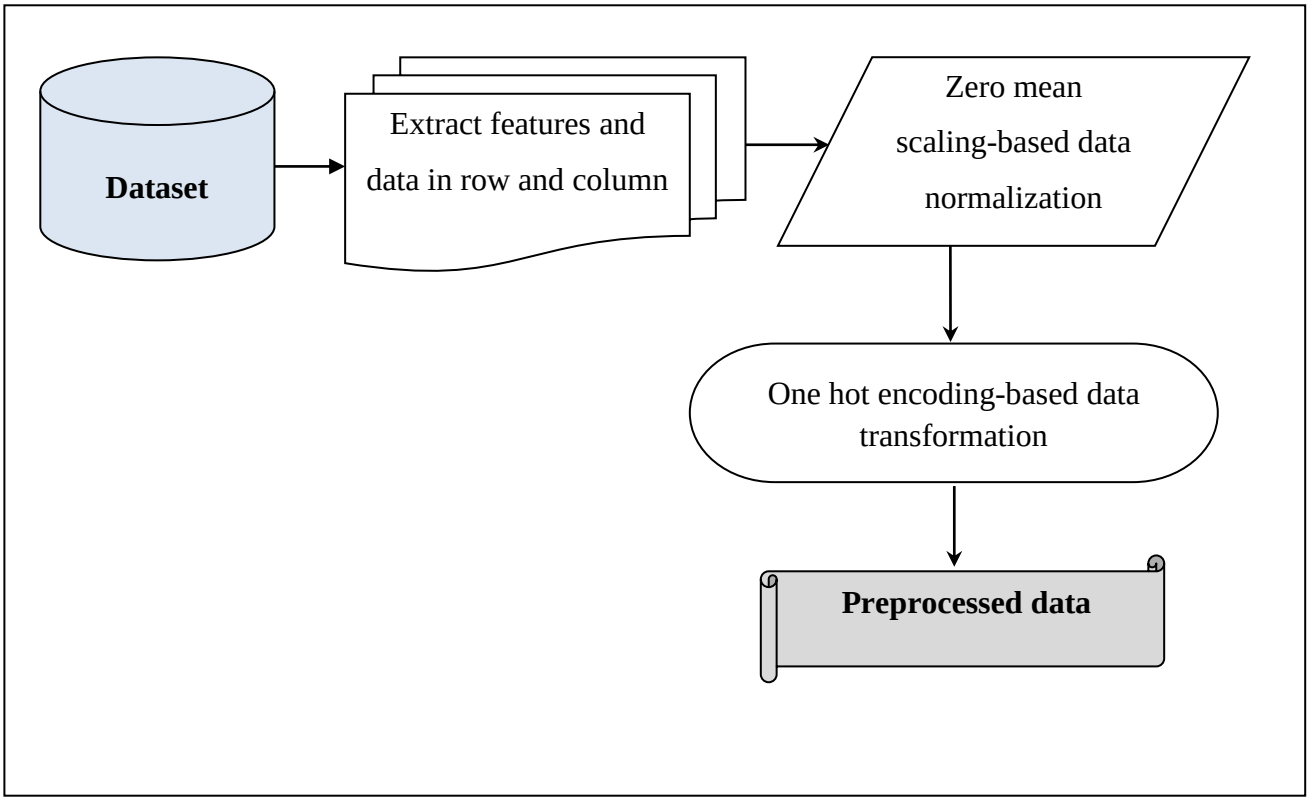


Figure 4.2 Zero Mean Feature Normalized Encoding based Preprocessing Model

Figure 4.2 demonstrates the pre-processing of the dataset using ZMFNE Technique. Initially, the sample input matrix is represented in both rows and columns, depending on the patient data and features. Then, data normalisation is carried out for each sample through zero-

mean feature scaling. Finally, the normalised data is transformed into a binary visualisation using the one-hot encoding technique.

Assume the attributes (i.e., features) '$a_j = a_1, a_2, .., a_m$' with the corresponding patient data samples $D_i = d_1, d_2, .., d_n$' in the input dataset. Then, the input sample matrix 'M' is expressed as given below.

$$M = \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} & \ldots. & \boldsymbol{a_m} \\ d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ d_{m1} & d_{m2} & \ldots & d_{mn} \end{bmatrix} \quad (4.6)$$

In above equation (4.6), sample matrix 'M' is formed through the number of attributes and data. In order to normalize data features, zero mean feature scaling method is applied and it is calculated by,

$$Fs = \left[\frac{a_i - M}{\sigma}\right] \quad (4.7)$$

Where 'Fs' indicates feature scaling, 'a_i' refers to an input attribute (feature), 'M' refers to the mean of the attribute, and 'σ' denotes a deviation.

Upon completing the feature scaling process, data transformation is carried out using one hot encoding the encoder modifies the categorical features into a numeric array. The input to this encoder is considered as the integer array forms such as array-like or strings and it is denoted by categorical features. Features are encoded as a binary column in terms of [0,1] for each category and gives a sparse matrix.

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \dots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} \quad (4.8)$$

Whereas 'B' gives the binary representation, and 'b' denotes an output in a matrix form. From this, the data normalization and transformation process is carried out, thereby achieving pre-processed output. The algorithm of data pre-processing using ZMFNE technique is described as follows.

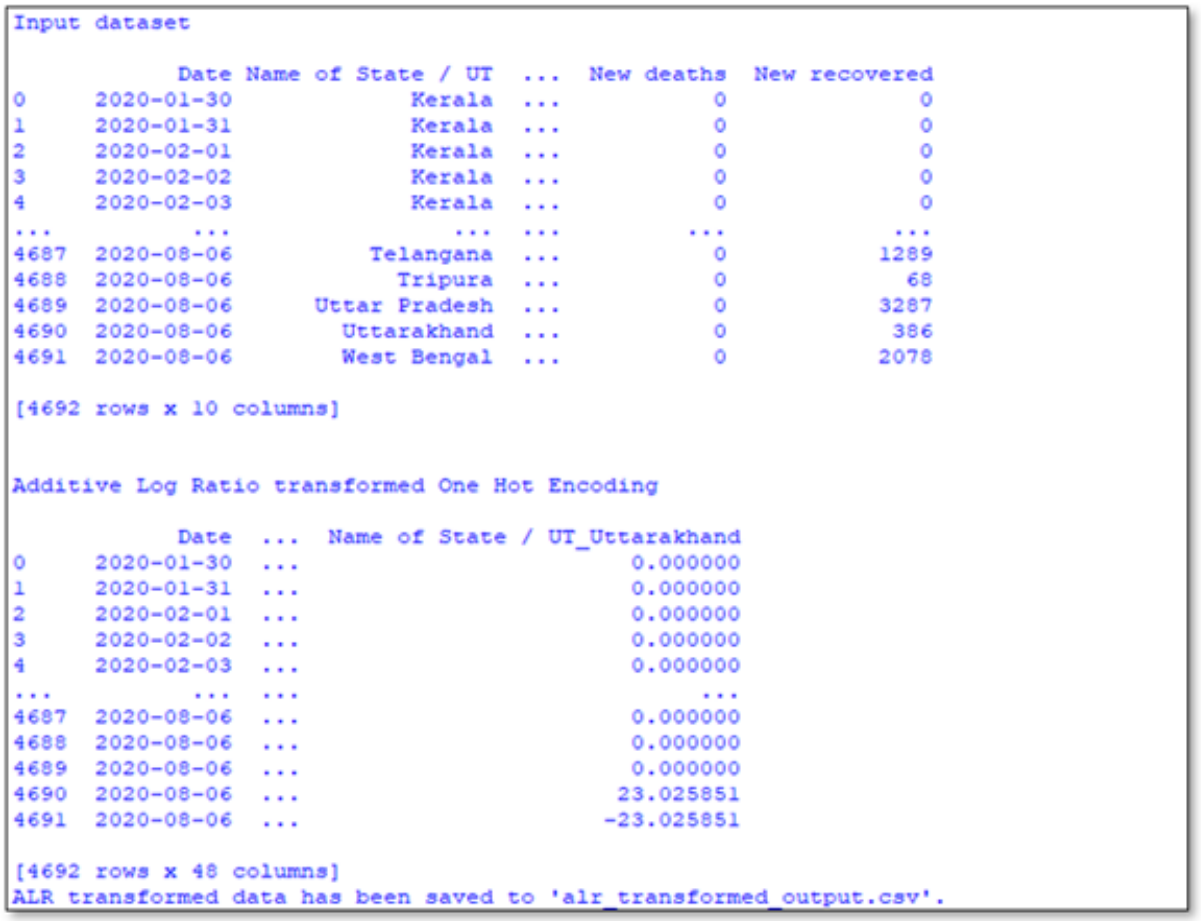
// **Algorithm 4.2 Zero mean feature normalized encoding-based Data Pre-processing**
Input: Dataset D, attributes $A = a_1\ a_2,..,a_m$'
Output: Obtain the preprocessed data
Begin **step 1:** Collect the number of attributes $A = a_1\ a_2,..,a_m$ and data from the dataset '$D_i = d_1, d_2,..,d_n$' **step 2:** **For** each attribute in the dataset **step 3**: Form the input sample matrix 'M' **step 4**: Apply zero mean feature scaling using (4.7) **step 5:** Normalize the data in the ranges from 0 to 1 **step 6:** **End for** **step 7:** Perform one hot encoding and obtain the binary representation of data '0' or '1' **End**

Algorithm 4.2 explains the process of data normalisation and encoding to obtain the preprocessed data output. This can be performed by initially collecting input data samples and features from the dataset. Then, zero-mean feature scaling is employed to normalise the input data. Next, one-hot encoding is employed to obtain the binary representation of the data. Data-preprocessed outcomes are achieved to decrease time and space complexity.

4.4 EXPERIMENTAL SETUP

The results of the above-mentioned techniques are examined with COVID-19 Corona Virus India Dataset and the RSNA Pneumonia Detection Challenge data repository. COVID-19 dataset and RSNA Pneumonia datasets are collected from the Kaggle dataset. The RSNA Pneumonia Detection Challenge dataset is used to detect pneumonia in medical images. In our work, the 10 cross-validation is employed. The entire data repository is separated into training and testing sets. Here, the majority of the information in the dataset, i.e., 80% of the data, is considered for training, and 20% of the data is reserved for testing. However, the 80:20 splitting ratio is most effective when applied to an input dataset, and its justification is derived from the Pareto principle. To conduct the above experiments, the mentioned files are considered for preprocessing the input dataset. The performance of proposed preprocessing techniques, such as the ALRTOHE technique and ZMFNE technique, was evaluated in comparison to an existing CNN-GRU-based hybrid deep learning method and a Variant of Concerned Deep Learning (VOC-DL) prediction framework. The outcome of both suggested and current methods are evaluated using the testing metrics listed below.

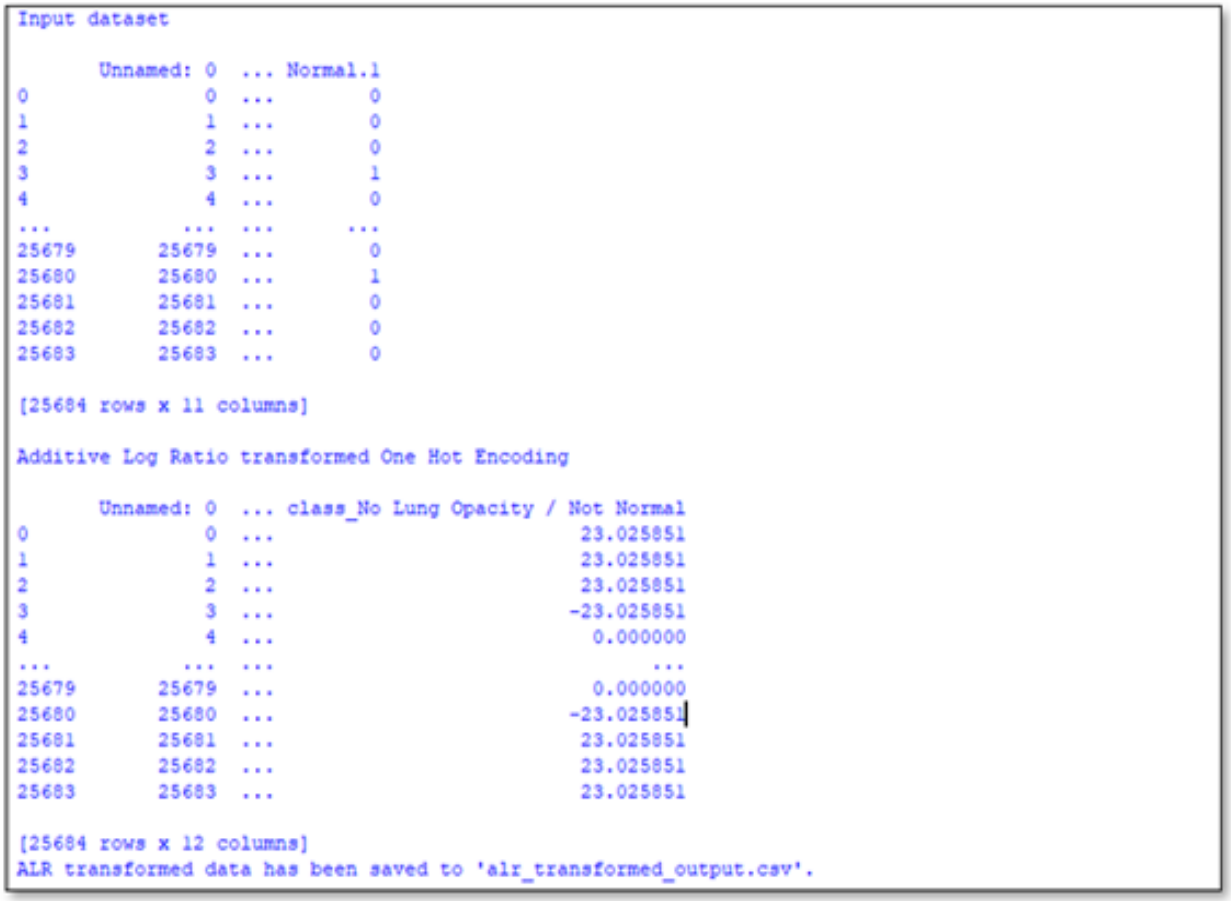
- Preprocessing accuracy
- Preprocessing time
- Space complexity
- Error rate



```
Input dataset

            Date Name of State / UT  ...  New deaths  New recovered
0     2020-01-30             Kerala  ...           0              0
1     2020-01-31             Kerala  ...           0              0
2     2020-02-01             Kerala  ...           0              0
3     2020-02-02             Kerala  ...           0              0
4     2020-02-03             Kerala  ...           0              0
...          ...                ...  ...         ...            ...
4687  2020-08-06          Telangana  ...           0           1289
4688  2020-08-06            Tripura  ...           0             68
4689  2020-08-06      Uttar Pradesh  ...           0           3287
4690  2020-08-06        Uttarakhand  ...           0            386
4691  2020-08-06        West Bengal  ...           0           2078

[4692 rows x 10 columns]


Additive Log Ratio transformed One Hot Encoding

            Date  ...  Name of State / UT_Uttarakhand
0     2020-01-30  ...                        0.000000
1     2020-01-31  ...                        0.000000
2     2020-02-01  ...                        0.000000
3     2020-02-02  ...                        0.000000
4     2020-02-03  ...                        0.000000
...          ...  ...                             ...
4687  2020-08-06  ...                        0.000000
4688  2020-08-06  ...                        0.000000
4689  2020-08-06  ...                        0.000000
4690  2020-08-06  ...                       23.025851
4691  2020-08-06  ...                      -23.025851

[4692 rows x 48 columns]
ALR transformed data has been saved to 'alr_transformed_output.csv'.
```

Figure 4.3 Results of ALRTOHE Preprocessing using COVID-19 Dataset

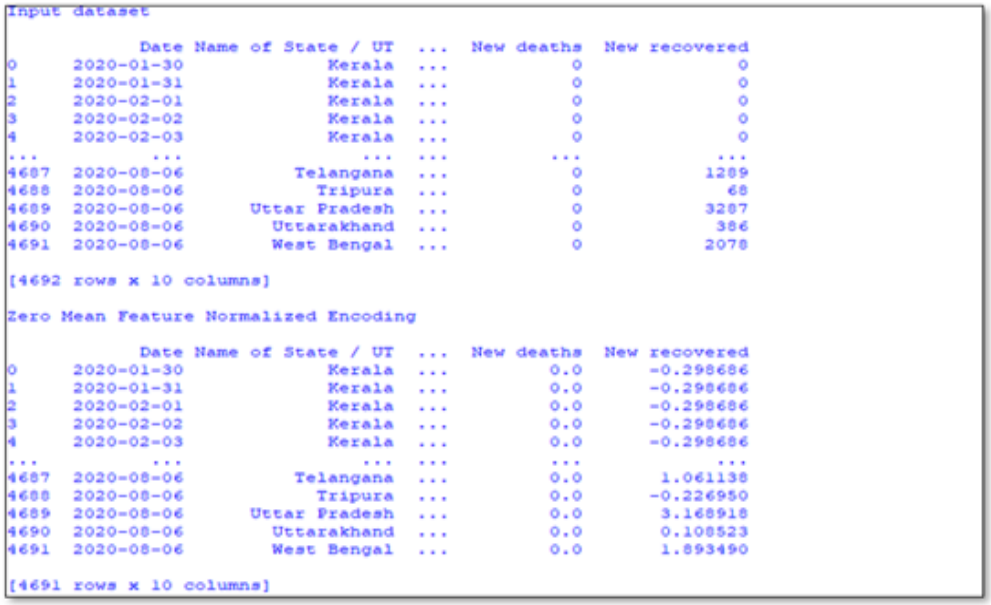
Figure 4.3 demonstrates outcomes attained using ALRTOHE Preprocessing for COVID-19 Dataset. As shown in the above Figure, the input dataset is the COVID-19 Dataset, which includes data, the name of the state, new deaths, and newly recovered information. To preprocess the input database, an Additive Log Ratio Transformed One Hot Encoding is applied. Here, the additive log-ratio transformation is used to normalise the information within a specific range. Next, decoding data is implemented to transfer numerical database within binary coding. Subsequently, One Hot Encoding converts the categorical information into a format suitable for machine learning methods. From where the preprocessed outcomes are obtained.



```
Input dataset

       Unnamed: 0  ... Normal.1
0               0  ...        0
1               1  ...        0
2               2  ...        0
3               3  ...        1
4               4  ...        0
...           ...  ...      ...
25679       25679  ...        0
25680       25680  ...        1
25681       25681  ...        0
25682       25682  ...        0
25683       25683  ...        0

[25684 rows x 11 columns]

Additive Log Ratio transformed One Hot Encoding

       Unnamed: 0  ... class_No Lung Opacity / Not Normal
0               0  ...                          23.025851
1               1  ...                          23.025851
2               2  ...                          23.025851
3               3  ...                         -23.025851
4               4  ...                           0.000000
...           ...  ...                                ...
25679       25679  ...                           0.000000
25680       25680  ...                         -23.025851
25681       25681  ...                          23.025851
25682       25682  ...                          23.025851
25683       25683  ...                          23.025851

[25684 rows x 12 columns]
ALR transformed data has been saved to 'alr_transformed_output.csv'.
```

Figure 4.4 Results of ALRTOHE Preprocessing using Pneumonia dataset

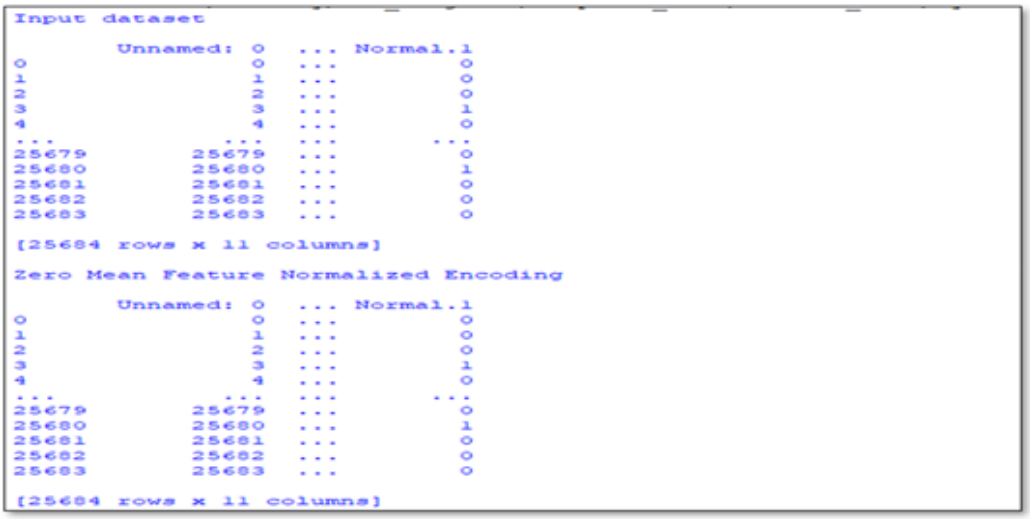
Data preprocessing using the ALRTOHE model for the Pneumonia dataset is shown in Figure 4.4. Input data is taken from the Pneumonia Dataset. The database contains unnamed columns, as well as patient ID, target, class, normal, lung opacity, and other relevant information. By applying ALRTOHE, using One Hot Encoding categorical data are converted into a numerical format, which is effectively utilised through ML techniques. From this, the dimensionality of both datasets was reduced while obtaining the preprocessed outputs.



```
Input dataset

            Date Name of State / UT  ...  New deaths  New recovered
0     2020-01-30              Kerala  ...           0              0
1     2020-01-31              Kerala  ...           0              0
2     2020-02-01              Kerala  ...           0              0
3     2020-02-02              Kerala  ...           0              0
4     2020-02-03              Kerala  ...           0              0
...          ...                 ...  ...         ...            ...
4687  2020-08-06           Telangana  ...           0           1289
4688  2020-08-06             Tripura  ...           0             68
4689  2020-08-06       Uttar Pradesh  ...           0           3287
4690  2020-08-06         Uttarakhand  ...           0            386
4691  2020-08-06         West Bengal  ...           0           2078

[4692 rows x 10 columns]

Zero Mean Feature Normalized Encoding

            Date Name of State / UT  ...  New deaths  New recovered
0     2020-01-30              Kerala  ...         0.0      -0.298686
1     2020-01-31              Kerala  ...         0.0      -0.298686
2     2020-02-01              Kerala  ...         0.0      -0.298686
3     2020-02-02              Kerala  ...         0.0      -0.298686
4     2020-02-03              Kerala  ...         0.0      -0.298686
...          ...                 ...  ...         ...            ...
4687  2020-08-06           Telangana  ...         0.0       1.061138
4688  2020-08-06             Tripura  ...         0.0      -0.226950
4689  2020-08-06       Uttar Pradesh  ...         0.0       3.168918
4690  2020-08-06         Uttarakhand  ...         0.0       0.108523
4691  2020-08-06         West Bengal  ...         0.0       1.893490

[4691 rows x 10 columns]
```

Figure 4.5 Results of ZMFNE Preprocessing using COVID-19 Dataset

Figure 4.5 demonstrates the results generated for ZMFNE Preprocessing with COVID-19 Dataset. ZMFNE is a technique for standardising the features of a dataset to enhance disease prediction performance. It involves adjusting aspects so that they have a mean equals zero and a standard deviation equals one, making it easier for learning algorithms to converge and improve their performance. Additionally, one-hot encoding transforms categorical data into a binary format to minimise the complexity of dataset processing through machine learning (ML) techniques.



```
Input dataset

       Unnamed: 0  ...  Normal.1
0               0  ...         0
1               1  ...         0
2               2  ...         0
3               3  ...         1
4               4  ...         0
...           ...  ...       ...
25679       25679  ...         0
25680       25680  ...         1
25681       25681  ...         0
25682       25682  ...         0
25683       25683  ...         0

[25684 rows x 11 columns]

Zero Mean Feature Normalized Encoding

       Unnamed: 0  ...  Normal.1
0               0  ...         0
1               1  ...         0
2               2  ...         0
3               3  ...         1
4               4  ...         0
...           ...  ...       ...
25679       25679  ...         0
25680       25680  ...         1
25681       25681  ...         0
25682       25682  ...         0
25683       25683  ...         0

[25684 rows x 11 columns]
```

Figure 4.6 Results of ZMFNE Preprocessing using Pneumonia Dataset

The preprocessing results of Pneumonia Dataset using ZMFNE technique is shown in Figure 4.6. Zero mean feature normalization integrated with encoding ensures features are on the identical scale, and that categorical variables are changed into a numerical form. With this, the binary representation of data is acquired to solve the computational complexity involved in the dataset.

4.5 CHAPTER SUMMARY

The proposed Additive Log Ratio Transformed One Hot Encoding (ALRTOHE) and Zero Mean Feature Normalised Encoding (ZMFNE) Techniques are designed to preprocess the data repository for accurate disease prediction. First, the ALRTOHE technique is designed to reduce the time and space complexity involved in preprocessing. In the ALRTOHE technique, an additive log-ratio transformation is performed to obtain normalised information. Then, one-hot encoding is employed to convert numerical information into a base two representation. As a result, data preprocessing is employed to obtain the binary representation of information, thereby reducing space complexity. Additionally, the ZMFNE Technique has been developed for preprocessing the dataset. Here, the sample input matrix generated depended on the patient data and features, which were represented in rows and columns. Then, data normalization is performed for each sample with the assistance of zero mean feature scaling. Lastly, the normalised data are transformed into a base two representation using the one-hot encoding method. The preprocessed data is obtained for disease detection, and the Outcome of preprocessing accuracy using the ZMFNE Technique is improved in terms of minimum preprocessing time and space complexity compared to other methods. Experimental evaluation of two proposed preprocessing techniques with existing and proposed classifiers (briefly discussed in chapter 8) shows that proposed preprocessing techniques yield better outcome in terms of accuracy as 85 % , 87% for COVID-19 dataset, 89% and 90% for Pneumonia dataset, also other metrics time, space complexity and error rate are discussed in the upcoming chapters. The next chapter presents proposed feature selection techniques and the reduced features in detail.