CHAPTER 3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

The coronavirus has had a massive international impact, negatively affecting the day-to-day lives of millions of people. It remains a challenge to predict the emergence of cases, and many governments have struggled to comprehend the scale and impact of the virus. The exponential spread of the virus (including its variants) can be curbed only if the entire population is vaccinated or it has been eliminated from the population. It is a crucial determining factor for the early diagnosis of diseases like COVID-19 and Pneumonia. A novel proposed research work has been developed to achieve higher accuracy in disease prediction. To meet this need, this proposed research utilises preprocessing and feature selection to achieve results in a shorter time frame and reduce the error rate. Also, the classification is performed to forecast the normal and disease patient data with maximum accuracy.

3.2 EXISTING WORK

To forecast the spread of the epidemic, Anil Utku (2023) developed a CNN-GRU-based hybrid deep learning model. It failed to improve disease prediction accuracy. VOC-DL prediction framework was examined by Zhifang Liao *et al.*,(2022) for predicting daily new confirmed cases. However, the prediction time was higher.

Hybrid Chi^2-MI basis feature selection model was discussed by Samrat Kumar Dey *et al.*,(2022) for distinguishing between Chronic Kidney Disease and non- Chronic Kidney Disease. However, real-time diagnosis of kidney failure patients was not performed.

A deep learning method based on LSTM was developed by Shastri S *et al.*,(2021) to predict disease prediction. However, protective measures were not provided to predict the cases of illness. DSPM was examined by D Ayris *et al.*,(2022) with minimal error rates, and the performance rate with respect to disease prediction was not improved.

3.3 PROPOSED METHODOLOGY

The proposed research work employs the following techniques to accurately classify the data with high accuracy in minimum time, which is shown in Figure. 3.1.

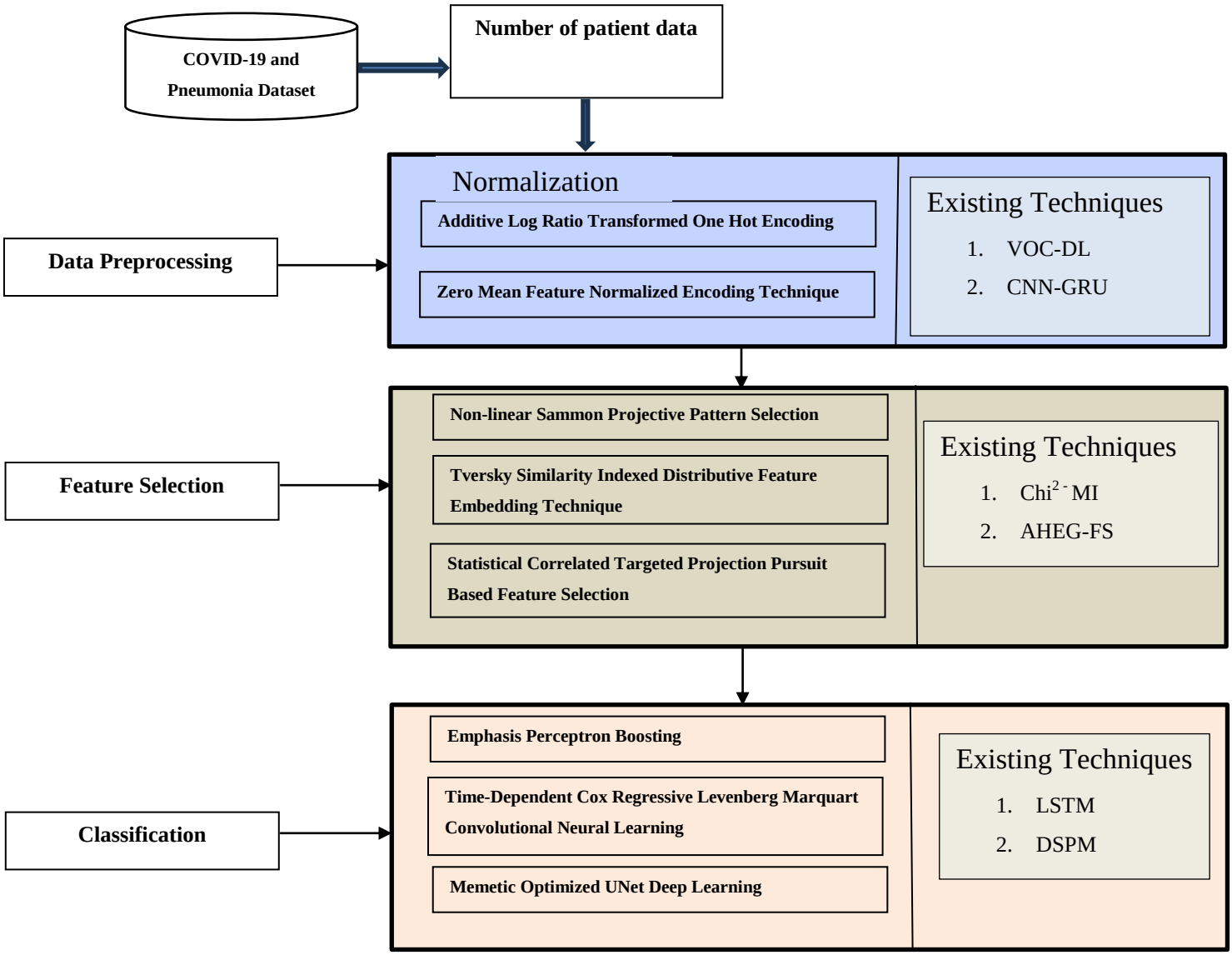

Figure.3.1 Overview of Proposed Methodology

3.3.1 Proposed Additive Log Ratio Transformed One Hot Encoding-Based Data Preprocessing

Initially, the proposed ALRTOHE technique is developed to eradicate the noise using additive log-ratio transformation and the one-hot encoding technique. Using additive log-ratio transformation, input data is considered. This aids in grouping data in a similar format, offering consistency for further processing. One-hot encoding is utilised in the ALRTOHE technique to convert numerical variables into binary vectors. As a result, the proposed ALRTOHE technique achieves preprocessing accuracy with reduced preprocessing time, space complexity, and error rate.

3.3.2 Proposed Zero Mean Feature Normalized Encoding Technique

The ZMFNE technique is developed to eliminate noisy data and dimensionality issues. The data is transformed into a matrix, and zero-mean feature scaling is used to normalise the data. The data transformation process is carried out with the aid of one-hot encoding to preprocess the data with minimal time and space complexity. The results show that the preprocessing accuracy of the ZMFNE technique is enhanced in minimal time and with a lower error rate.

3.3.3 Proposed Nonlinear Sammon Projective Pattern Selection

The proposed NSPPS technique is used to determine error-minimised optimal and pertinent patterns with improved accuracy. With the aid of Sammon projection, the data in the high-dimensional space is mapped to the low-dimensional space. Nonlinear Sammon Projection is used to identify applicable and unnecessary patterns. The error-reduced pertinent patterns are only taken for disease prediction with a shorter time. The NSPPS technique achieved higher feature selection accuracy with less time.

3.3.4 Proposed Tversky Similarity Indexed Distributive Feature Embedding Technique

The proposed TSIDFE technique is developed to perform appropriate feature selection with minimum error. The Tversky index of similarity between features is used to classify applicable and unrelated features. The correct identification is done with a similarity coefficient

where ‘0’ denoting the feature is considered unrelated and ‘1’ representing the feature is pertinent. Relevant features are selected with less feature selection time to obtain maximum feature selection accuracy.

3.3.5 Proposed Statistical Correlative Targeted Projection Pursuit-Based Feature Selection

The proposed SCTPP-FS technique is employed to decide the vital features for disease prediction. Kaiser–Meyer–Olkin selects target features based on correlative projection pursuit. The correlation between the features is measured to find the appropriate and immaterial features. The most significant features employed for accurate disease prediction are chosen. The feature selection accuracy is enhanced by reducing both feature selection time and error rate.

3.3.6 Proposed Emphasis Perceptron Boosting Classification

The proposed EPBC utilises a perceptron binary classifier to categorise the input data through a weighted sum. Emphasis boosting is constructed during the weighted sum that divides the patterns or data by means of zero training error. In this manner, precise prediction is obtained with enhanced precision, recall, specificity and lesser prediction time.

3.3.7 Proposed Time-Dependent Cox Regressive Levenberg–Marquardt Convolutional Neural Learning Technique

The proposed TCLMCNL technique is developed to execute data classification for disease forecasting. Various kinds of layers are employed in TCLMCNL. The time-dependent Cox regression is used to measure the relationship between the data using Cramér's phi correlation function. Additionally, the Huber loss is employed to provide error classification results. The results of TCLMCNL show an increase in prediction accuracy and precision with reduced prediction time.

3.3.8 Proposed Memetic Optimized U-Net Deep Learning Classifier

The proposed MO-UNetDL technique is introduced with higher performance of disease prediction during classification. Wilcox's index coefficient discovers the similarity among input data. Then, the max-pooling operation is carried out to diminish the data dimension. Finally, data classification outcomes are obtained. Memetic optimisation is used to tune hyperparameters to

obtain optimised classification results. Therefore, the result of prediction accuracy and recall is enhanced within less time for prediction.

3.4. EXPERIMENTAL SETUP

The performance outcomes of proposed techniques such as ALRTOHE, ZMFNE, NSPPS Model, TSIDFE, SCTPP-FS, EPBC, TCLMCNL and MO-UNetDL are implemented in Python.

The COVID-19 database in Kaggle consist of eight.csv files. COVID-19 dataset is gathered from https://www.kaggle.com/datasets/imdevskp/COVID19-corona-virus-india-dataset. This database presents daily case reports, USA daily state reports and Time series summaries. Daily case reports comprise of attributes such as FIPS, Admin2 and others. RSNA Pneumonia Recognition Challenge database is employed to discover pneumonia in medical images. The dataset is collected from https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. Sample dataset used for this work is provided in [Annexure]. Over 5-years-old children have 15% of Pneumonia deaths globally. In United States, pneumonia accounts for 500,000 patience admission to emergency units and 50,000 accounting to deaths.

3.5. PERFORMANCE METRIC ANALYSIS

The performance evaluation of the proposed techniques is developed for disease prediction in the COVID-19 coronavirus dataset and RSNA Pneumonia Detection Challenge dataset using following metrics.

- Preprocessing accuracy
- Preprocessing Time (PT)
- Error Rate (ER)
- Space Complexity (SC)
- Feature selection accuracy
- Feature selection time (FT)
- Prediction accuracy
- Recall
- F-measure

- Prediction time (PT)

3.5.1 Preprocessing accuracy

It is measured as the ratio of number of properly preprocessed data samples to the total number of data samples.

$$P_{Acc}=(Nd_{cp}/Td)*100 \qquad (3.1)$$

Where, 'P_{Acc}' is preprocessing accuracy, 'Nd_{cp}' is denoted as number of data accurately preprocessed and 'Td' represents total amount of data.

3.5.2 Preprocessing Time

It is defined as the time consumed for preprocessing.

$$P_{time} = \sum_{i=1}^{n} n * Time\ [psd] \qquad (3.2)$$

From equation (3.2), 'P_{time}' is preprocessing time, 'n' represents number of sample data. '$\boldsymbol{Time}[\boldsymbol{psd}]$' is sample data. It is determined by milliseconds (ms).

3.5.3 Space complexity

It is measured as the memory space taken for data preprocessing.

$$S_c= \sum_{i=1}^{n} n * Mem_{psd} \qquad (3.3)$$

Where, 'S_c' is space complexity, and 'Mem_{psd}' indicates number of patient data samples from dataset. It is measured in Megabytes (KB).

3.5.4 Performance of Feature Selection Accuracy

It is determined as the ratio of number of data feature accurately chosen.

$$FS_{acc} = \frac{Ndf_{cs}}{T_{df}} * 100 \qquad (3.4)$$

Where 'FS_{acc}' denotes Feature Selection accuracy, 'Ndf_{cs}' indicates the number of data feature properly selected and 'T_{df}' total number of data features.

3.5.5 Feature Selection Time

It is discovered as the amount of time consumed to choose the pertinent feature data.

$$F_{st} = \sum_{i=1}^{n} Ndf_I * \text{Time [sdf]} \quad (3.5)$$

Where 'F_{st}' is feature selection time, 'Ndf_i' denotes number of data feature samples and '$Time[sdf]$' is a time for solo data feature selection. It is computed by milliseconds (ms).

3.5.6 Prediction Accuracy

It is defined as the proportion of patient data samples that are precisely predicted as normal or confirmed disease cases during classification.

$$PAC = \frac{tp+tn}{tp+tn+f_p+f_n} *100 \quad (3.6)$$

Where 'PAC' is prediction accuracy, 'tp' indicates true positive, 'tn' denotes true negative, 'f_p' is false positive, 'f_n' indicates false negative. The accuracy metric is determined in percentage (%).

3.5.7 Precision

Precision is measured based on true positives and false negatives.

$$Pre = \frac{tp}{tp+fp} *100 \quad (3.7)$$

Where 'Pre' is Precision and estimated in percentage (%).

3.5.8 Recall

It is computed based on both true positives as well as false negatives *Chapter 3* prediction.

$$\text{Rec} = \frac{tp}{tp+fn} * 100 \quad (3.8)$$

Where 'Rec' refers a recall and is computed in percentage (%).

3.5.9 Specificity

It is determined on the basis of true negatives as well as false positives of disease prediction.

$$\text{Spec} = \frac{tn}{tn+fp} * 100 \quad (3.9)$$

Where '$Spec$' is specificity and computed in percentage (%).

3.5.10 Prediction time

It is measured as the time taken by the algorithm to predict the patient's outcome through classification process.

$$\text{PTime} = \sum_{i=1}^{n} P_i * \text{Time [PSP]} \quad (3.10)$$

Where '$PTime$' indicates prediction time, 'P_i' denotes patients, '$Time[PSP]$' denotes time consumed to categorize the sole patient data sample. It is measured in milliseconds (ms).

3.5.11 F-measure

F-measure is measured based on precision and recall results. F-measure is mathematically formulated below,

$$\text{F} - \text{measure} = \left[2 * \frac{P_r * R_c}{P_r + R_c}\right] * 100 \quad (3.11)$$

Where, '$F - measure$' is computed based on precision P_r and recall 'R_c'. It is measured in terms of percentage (%).

3.5.12 Percentage Improvement

$$\text{Percentage Improvement} = \frac{\text{Proposed Value} - \text{Existing Value}}{\text{Existing Value}} * 100 \qquad (3.12)$$

The percentage improvement is calculated by dividing the difference between the proposed and existing values by the existing value, and multiplying the result by 100.

3.6. CHAPTER SUMMARY

The chapter discusses the research methodology employed for accurate disease prediction. This research is designed in three phases: preprocessing, feature selection, and classification, which are integral to the methodology. Thus, the proposed method is evaluated based on the following metrics: preprocessing accuracy, preprocessing time, error rate, space complexity, feature selection accuracy, feature selection time, prediction accuracy, recall, F-measure, and prediction time. The overall explanation for the design of the proposed work, which aims to improve accuracy with minimal time, is also briefly outlined in this chapter. The next chapter will describe the proposed preprocessing techniques and its results.