CHAPTER 8

INVESTIGATION ON NOVEL PREPROCESSING, FEATURE SELECTION, CLASSIFICATION TECHNIQUES USING DISEASE DATASET

8.1 INTRODUCTION

A wide range of medical diagnoses relies on examining disease data obtained from various datasets. The relevance of AI in evaluating medical data has enabled precise assessments to be carried out robotically. This minimized the task for doctors and the time required for analysis, improving the performance of disease prediction. AI helps identify environmental regions where disease or high-risk behaviors are prevalent. Machine learning (ML) and deep learning (DL) are two key sub-areas in advanced AI techniques. Machine learning (ML) and deep learning (DL) have been effectively used to enhance the accuracy of computer-aided diagnosis systems. Machine learning methods have achieved notable success in medical studies due to advanced methods that facilitate the automated extraction of relevant aspects. DL methods provide promising outcomes in medical study, namely fusion, recording, and classification.

Before classifying medical data, preprocessing is essential to eliminate noise and normalize data to improve integrity. By applying a preprocessing process, redundancy reduced structured database is obtained. Additionally, feature selection involves identifying the most relevant aspects of a given dataset. The primary main objective of feature selection is to increase accuracy by identifying a minimal subset of features that retains the comprehensive relevant details. For this purpose, proposed preprocessing methods such as the Additive Log Ratio Transformed One Hot Encoding (ALRTOHE) Technique, Zero Mean Feature Normalized Encoding (ZMFNE) technique, and feature selection methods such as Nonlinear Sammon Projective Pattern Selection (NSPPS) Model, Tversky Similarity-Indexed Distributive Feature Embedding (TSIDFE) Technique and Statistical correlative targeted projection pursuit-based feature selection (SCTPP-FS) Technique derived by ML and DL techniques are performance are experimentally assessed in this chapter. Then, we will analyze proposed classification methods,

such as the Emphasis Perceptron Boosting Classification (EPBC) method, the TCLMCNL Technique, and the Memetic Optimized U-Net Deep Learning (MO-UNetDL) method, using both with and without preprocessing methods. The benefits of using classification vary depending on the deep learning methods utilised for the disease dataset. However, the best model identified a combination of the MO-UNetDL method with the ZMFNE method, using SCTPP-FS, which achieved higher accuracy on the considered dataset.

8.2 PERFORMANCE ANALYSIS

Experimental evaluation of proposed two preprocessing models, three feature selection methods, and three classification methods are implemented in Python language. The outcomes of the initiated methods are examined using the RSNA Pneumonia Recognition Challenge Database https://www.kaggle.com/c/rsna-pneumonia-detection-challenge and COVID-19 database https://www.kaggle.com/datasets/imdevskp/covid19-corona-virus-india-dataset. The database is divided into distinct sets, including the training set and the testing set. The training set comprises 80% of patient data, and the testing set comprises 20% of data. The method mechanically detects the lung opacities on the chest radiographs.

8.2.1 Results of Preprocessing Methods

The preprocessing methods, such as the ALRTOHE technique and ZMFNE technique using MO-UNetDL classifier, are compared with an existing CNN-GRU-based hybrid deep learning (DL) method and VOC-DL methods to assess the efficacy of the approach. Both proposed and existing methods are analyzed using metrics that include preprocessing accuracy, preprocessing time, error rate, and space complexity.

Table 8.1 Preprocessing Accuracy for COVID-19 and Pneumonia Datasets

Data Samples		Preprocessing Accuracy (%)							
		COVID-19 Dataset				Pneumonia Dataset			
COVID-19 Dataset	Pneumonia Dataset	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE
10000	2000	76.14	81.51	90.15	92.25	78.24	83.25	92.41	94.36
20000	4000	75.52	82.52	89.25	91.16	77.52	82.14	91.25	93.52
30000	6000	74.25	79.25	88.01	90.22	76.14	81.63	90.63	92.41
40000	8000	73.14	78.41	87.25	89.25	75.25	80.52	89.41	91.25
50000	10000	72.63	77.52	86.62	88.14	74.14	79.41	88.14	90.36
60000	12000	71.25	76.62	85.41	87.54	73.14	78.25	87.25	89.14
70000	14000	70.25	75.41	84.25	86.82	72	77.63	86.41	88.25
80000	16000	69.25	74.25	83.41	85.46	71.25	76.58	85.36	87.63
90000	18000	68.14	73.41	82.15	84.25	70	75.41	84.14	86.25
100000	20000	67.63	72.25	81.52	83.14	69.52	74.36	83.11	85.14

Table 8.1 presents a comparative analysis of preprocessing accuracy for the current and suggested methods utilising the Pneumonia dataset and the COVID-19 dataset. The results of preprocessing accuracy using the proposed ALRTOHE technique and the ZMFNE technique are compared with those of an existing CNN-GRU-based hybrid deep learning (DL) method and VOC-DL models. By analyzing the above tables, the preprocessing accuracy of the proposed ZMFNE technique is found to be superior to that of the other methods in both datasets.

This is due to the application of one hot encoding technique in the ZMFNE method. Zero-mean feature scaling is employed to normalize the samples. Normalized data is forwarded to the binary depiction using one hot encoding technique for correctly preprocessing the input samples. With this, the preprocessing accuracy is improved by 11% and 16% in the proposed ZMFNE technique compared to the CNN-GRU-based hybrid DL method and VOC-DL models, respectively, for the COVID-19 dataset. Similarly, the preprocessing accuracy is increased by

11% and 16% for the proposed ZMFNE technique compared to the CNN-GRU-based hybrid DL method and VOC-DL models for the Pneumonia dataset.

Table 8.2 Preprocessing Time for COVID-19 and Pneumonia datasets

Data Samples		Preprocessing Time(ms)							
		COVID-19 Dataset				Pneumonia Dataset			
COVID-19 Dataset	Pneumonia Dataset	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE
10000	2000	3900	3600	2400	2200	3700	3400	2200	2000
20000	4000	4200	3800	3000	2800	4000	3600	2800	2600
30000	6000	4500	4100	3400	3200	4300	3900	3200	3000
40000	8000	4800	4300	3800	3500	4600	4100	3600	3300
50000	10000	5100	4600	4000	3600	4900	4400	3800	3400
60000	12000	5400	4800	4200	3900	5300	4600	4000	3700
70000	14000	5700	5100	4500	4300	5500	4900	4300	4100
80000	16000	5900	5400	4800	4600	5700	5200	4600	4400
90000	18000	6200	5800	5300	5000	6000	5600	5100	4800
100000	20000	6400	6200	5500	5200	6200	6000	5300	5000

Table 8.2 presents a comparative study of preprocessing time based on patient data samples for various methods using the Pneumonia dataset and the COVID-19 dataset. As provided in the above Figure, the increasing tendency is examined when increasing samples. However, the proposed ZMFNE technique takes less time for preprocessing the patient data samples using both datasets.

To compare the ALRTOHE technique and ZMFNE technique with the existing CNN-GRU-based hybrid deep learning method and VOC-DL models, the validation of the method is carried out. From the above tables, it is evident that the preprocessing time of the ZMFNE technique is found to be lower than the other methods. The reason for the shorter preprocessing time is the application of zero-mean feature scaling. The data samples are normalized to obtain a structured form of data in the ZMFNE technique. Additionally, the binary representation of data

via one-hot encoding significantly reduces the time required to preprocess the input samples. Thus, the results of preprocessing time using the proposed ZMFNE technique are minimized by 22% and 29% compared to the CNN-GRU-based hybrid DL method and VOC-DL models, respectively, for the Pneumonia dataset. Additionally, the preprocessing time of the proposed ZMFNE technique is reduced by 21% and 27% compared to the CNN-GRU-based hybrid DL method and VOC-DL models, respectively, for the COVID-19 dataset.

Table 8.3 Space Complexity for COVID-19 and Pneumonia datasets

Data Samples		Space Complexity (KB)							
		COVID-19 Dataset				Pneumonia Dataset			
COVID-19 Dataset	Pneumonia Dataset	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE
10000	2000	53	48	41	36	51	46	39	34
20000	4000	56	51	44	39	54	48	42	37
30000	6000	59	53	47	42	57	51	45	40
40000	8000	62	56	51	45	60	54	48	42
50000	10000	65	61	55	46	62	59	52	44
60000	12000	67	65	58	49	65	62	56	46
70000	14000	70	67	62	51	68	65	60	49
80000	16000	72	70	65	55	70	68	63	52
90000	18000	75	72	68	57	73	70	66	55
100000	20000	77	75	72	61	75	73	70	59

Table 8.3 demonstrates the comparative results of space complexity using proposed and existing methods based on the patient data samples for two disease datasets. By observing the above table, the space complexity of the proposed ZMFNE technique is measured to be lesser than the other methods. The graphical illustration of space complexity using two proposed and two existing methods is shown below.

The validation of the proposed ALRTOHE technique and ZMFNE technique is carried out against existing CNN-GRU-based hybrid Deep Learning Methods and VOC-DL models. Ten

different runs are performed with various ranges of patient data samples. The results from the Figure confirm that the ZMFNE technique achieved lesser space complexity than the other proposed and existing works. The smaller space complexity is achieved by normalizing the input samples using zero-mean feature scaling, which enables obtaining a binary representation for understanding and processing the data. As a result, the space complexity is decreased by 14% and 18% in the proposed ZMFNE method using the COVID-19 dataset compared to the existing CNN-GRU-based hybrid method and VOC-DL models, respectively. In addition, the space complexity is minimized in the proposed ZMFNE method using the Pneumonia dataset by 14% and 18% as compared to the existing CNN-GRU-based hybrid DL method and VOC-DL models, respectively, for the COVID-19 dataset.

Table 8.4 shows the error rate for both datasets. The error rate is increased through enhancements in data samples. However, a comparatively smaller error rate is observed for the proposed methods. Among the two proposed methods, the proposed ZMFNE technique achieved the lowest error rate throughout preprocessing of the given data samples.

Table 8.4 Error Rate for COVID-19 and Pneumonia datasets

Data Samples		Error Rate (%)							
		COVID-19 Dataset				Pneumonia Dataset			
COVID-19 Dataset	Pneumonia Dataset	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE	Existing VOC-DL	Existing CNN	Proposed ALRTOHE	Proposed ZMFNE
10000	2000	23.86	18.49	9.85	7.75	21.76	16.75	7.59	5.64
20000	4000	24.48	17.48	10.75	8.84	22.48	17.86	8.75	6.48
30000	6000	25.75	20.75	11.99	9.78	23.86	18.37	9.37	7.59
40000	8000	26.86	21.59	12.75	10.75	24.75	19.48	10.59	8.75
50000	10000	27.37	22.48	13.38	11.86	25.86	20.59	11.86	9.64
60000	12000	28.75	23.38	14.59	12.46	26.86	21.75	12.75	10.86
70000	14000	29.75	24.59	15.75	13.18	28	22.37	13.59	11.75
80000	16000	30.75	25.75	16.59	14.54	28.75	23.42	14.64	12.37

90000	18000	31.86	26.59	17.85	15.75	30	24.59	15.86	13.75
100000	20000	32.37	27.75	18.48	16.86	30.48	25.64	16.89	14.86

In comparison to the conventional CNN-GRU-based hybrid method and VOC-DL models, the ALRTOHE and ZMFNE techniques achieve the lowest error rate. Particularly, the ZMFNE technique yields a lower error during preprocessing compared to the ALRTOHE technique. This is due to the zero-mean scaling-based normalization process and one-hot encoding data transformation. With this, data preprocessing is performed with lower error. Therefore, From the above Tables 8.1 to 8.4, the error rate of the ZMFNE technique is reduced by 47% and 57% compared to the conventional CNN-GRU-based hybrid deep learning method and VOC-DL models, respectively, for the COVID-19 dataset. Additionally the error rate of the ZMFNE technique is minimized by 53% and 62% compared to the conventional CNN-GRU-based hybrid DL method and VOC-DL models, respectively, using the Pneumonia dataset.

8.2.2 Results of Feature Selection Methods

Feature selection methods, such as the NSPPS method, TSIDFE method, and SCTPP-FS method, are compared with existing Chi^2-MI and AHEG-FS to evaluate the execution of the proposed techniques. The outcomes of proposed and existing techniques are examined using feature selection accuracy, feature selection time, space complexity, and error rate parameters.

Table 8.5 Feature selection results for COVID-19 and Pneumonia datasets

Models	COVID -19				Pneumonia			
	FSA (%)	FST (ms)	ER (%)	SC (KB)	FSA (%)	FST (ms)	ER (%)	SC (KB)
Existing Chi^2-MI	81	6900	19	76	84	6870	17	72
Existing AHEG-FS	84	6400	16	70	87	6360	15	65
Proposed NSPPS	86	5800	14	66	89	5760	12	60
Proposed TSIDFE	89	5320	11	61	91	5260	9	55
Proposed SCTPP-FS	**91**	**4950**	**9**	**55**	**93**	**4900**	**7**	**50**

Table 8.5 illustrates the performance analysis of feature selection accuracy with proposed preprocessing model for the NSPPS, TSIDFE and SCTPP-FS methods, as well as Chi^2-MI, and AHEG-FS, on both the COVID-19 and Pneumonia datasets. To show the efficacy of the proposed feature selection methods, a comparison is made with existing works. By examining the Table 8.5, it is evident that feature selection accuracy decreases as the number of patient data samples increases. However, comparatively, the feature selection accuracy of the third proposed SCTPP-FS method is increased than that of the alternative methods.

Compared to the preprocessing model, the feature selection accuracy of the SCTPP-FS method is improved over the other methods, namely the NSPPS method, TSIDFE method, Chi^2-MI, and AHEG-FS, when using the preprocessing model. Additionally, the feature selection time, error rate, and space complexity of the SCTPP-FS method are minimized when using the preprocessing model compared to conventional methods. The main reason behind the higher accuracy and lower time, error, and space complexity in feature selection is achieved in the proposed SCTPP-FS method through the use of Kaiser–Meyer–Olkin correlative projection pursuit. In this correlation measure, the relationship between features is calculated to identify pertinent and redundant features. Pertinent features are further used for the prediction process. This aids in improving feature selection accuracy and reducing time, error, and space complexity in the proposed SCTPP-FS.

The feature selection accuracy of the proposed SCTPP-FS is improved by 7% and 10% than the AHEG-FS and Chi^2-MI, respectively, for the COVID-19 dataset. Additionally, the feature selection accuracy of the SCTPP-FS method increases by 8% and 10% compared to the AHEG-FS and Chi^2-MI, respectively, for the Pneumonia dataset. The results of feature selection time using the SCTPP-FS method with the preprocessing model are decreased by 1450 ms and 1950 ms compared to AHEG-FS and Chi^2-MI, respectively, for the COVID-19 dataset.

With this, the space complexity of the proposed SCTPP-FS method with the preprocessing model is reduced by 21KB and 15KB than the Chi^2-MI and AHEG-FS, respectively, for the COVID-19 dataset. Likewise, the space complexity of the SCTPP-FS method without a preprocessing model is decreased by 22 KB and 15 KB than the existing methods for the Pneumonia dataset. The effects of the error rate utilising the proposed SCTPP-

FS method with a preprocessing model for the COVID-19 dataset are minimized as 9% and for the Pneumonia dataset error rate is decreased by 7% compared to the existing methods.

8.2.3 Results of Classification Methods

Here, the results of classification methods, including the EPBC technique, TCLMCNL technique, UNetDL technique, deep learning LSTM ensemble model, and DSPM, are analyzed with and without preprocessing methods. The result parameters considered to estimate the results of the above-mentioned classification methods are accuracy, precision, recall, and prediction time.

To confirm the efficiency of the proposed classification techniques, the comparison is made through two existing methods. From the simulation values, it is examined which accuracy is improved using three methods with preprocessing compared to the same methods without preprocessing. Compared to the three proposed methods, the MO-UNetDL technique achieves higher accuracy.

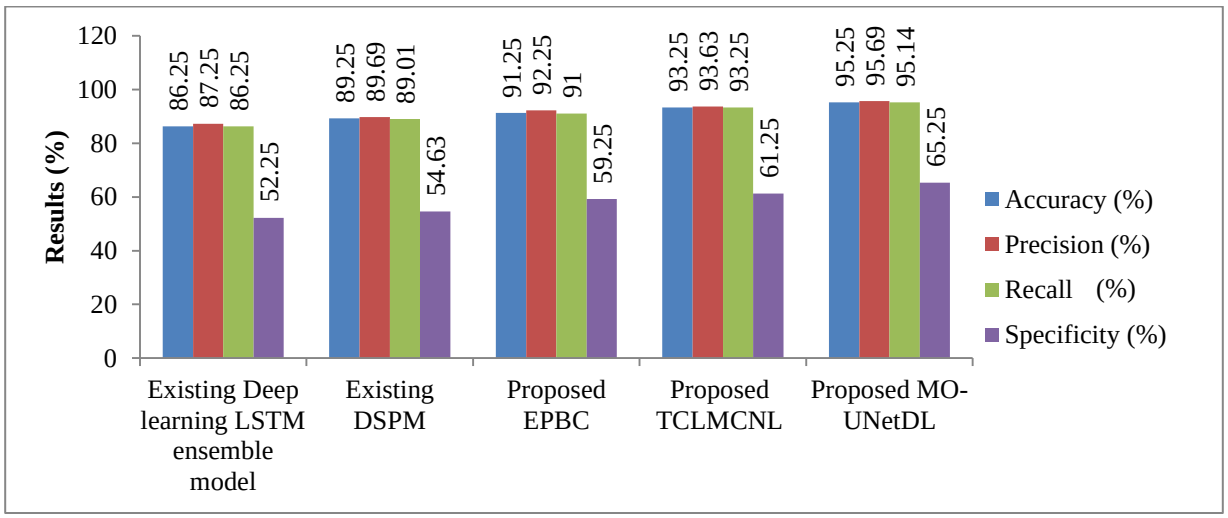

Figure 8.1 (a) Performance of Classification with preprocessing (ZMFNE) model for COVID-19 dataset

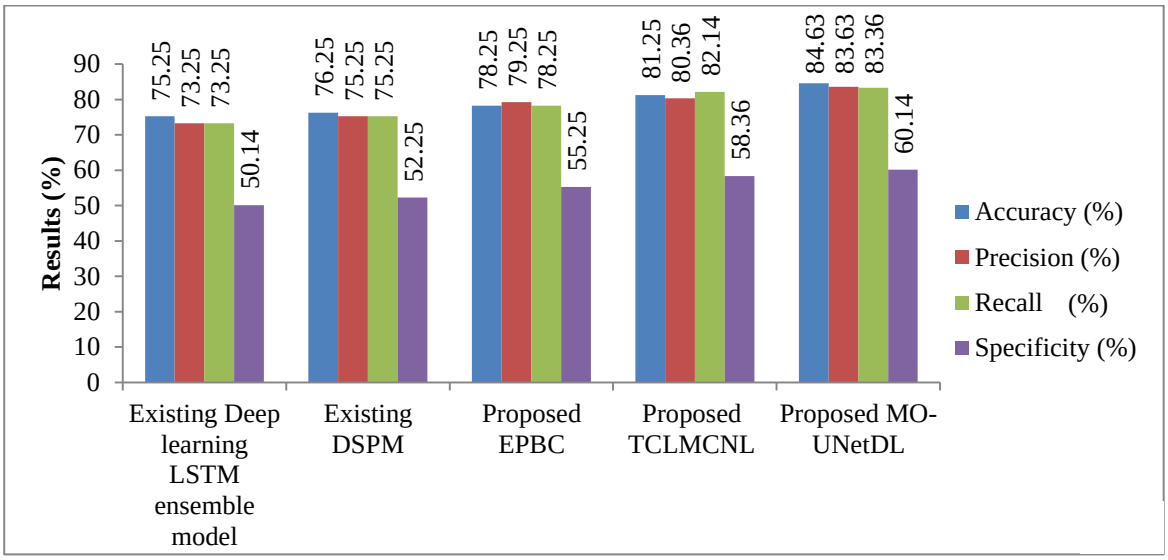

Figure 8.1 (b) Performance of Classification without preprocessing model for COVID-19 dataset

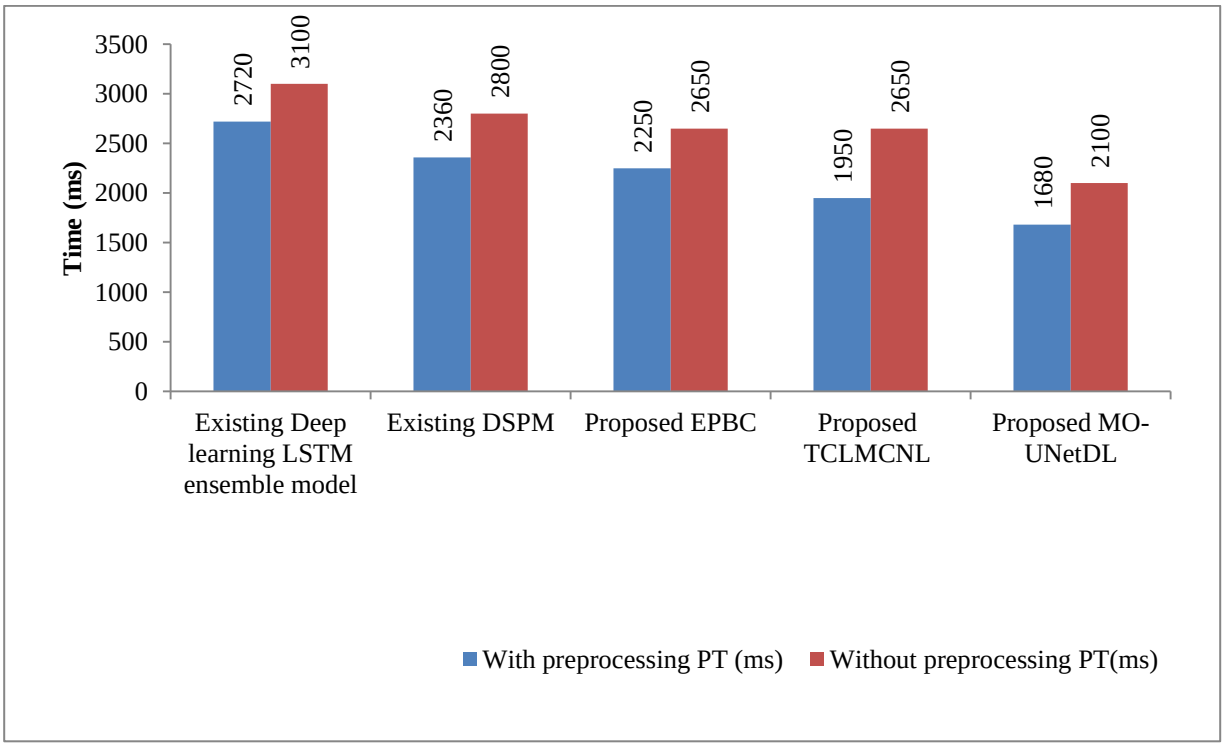

Figure 8.1 (c) Performance of prediction time with and without preprocessing model for COVID-19 dataset

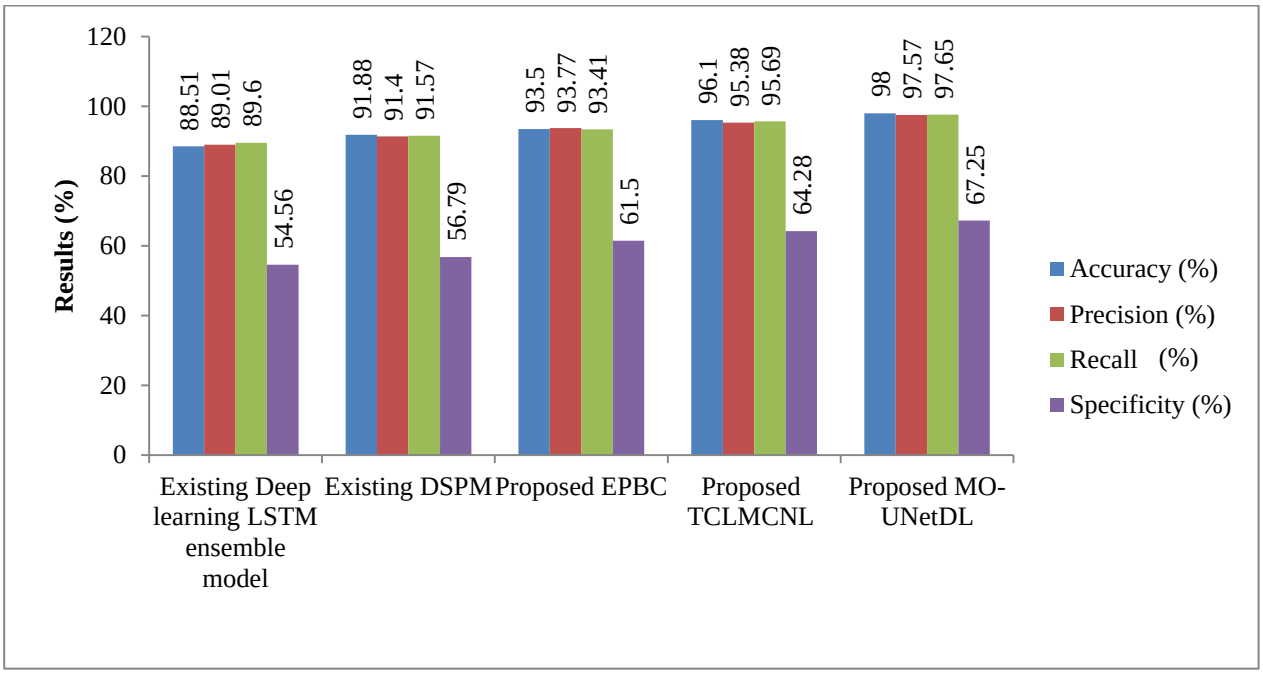

Figure 8.1 (d) Performance of Classification with Preprocessing (ZMFNE) model for Pneumonia dataset

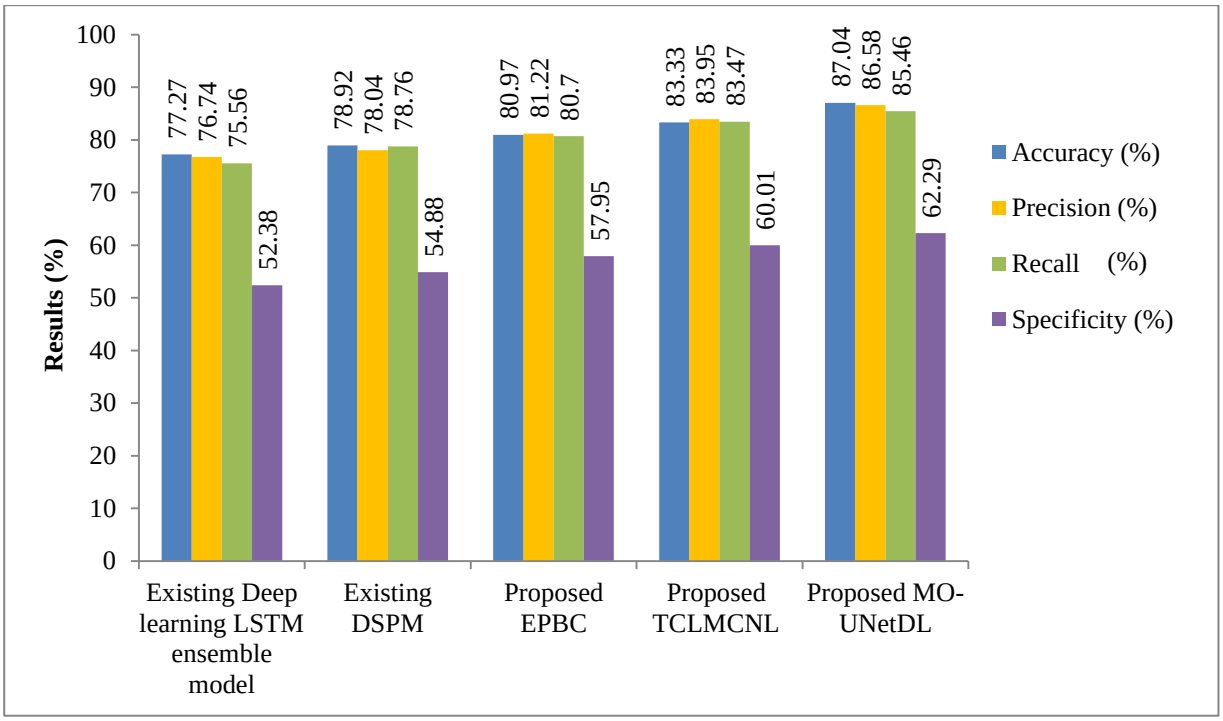

Figure 8.1(e) Performance of Classification without preprocessing model for Pneumonia dataset

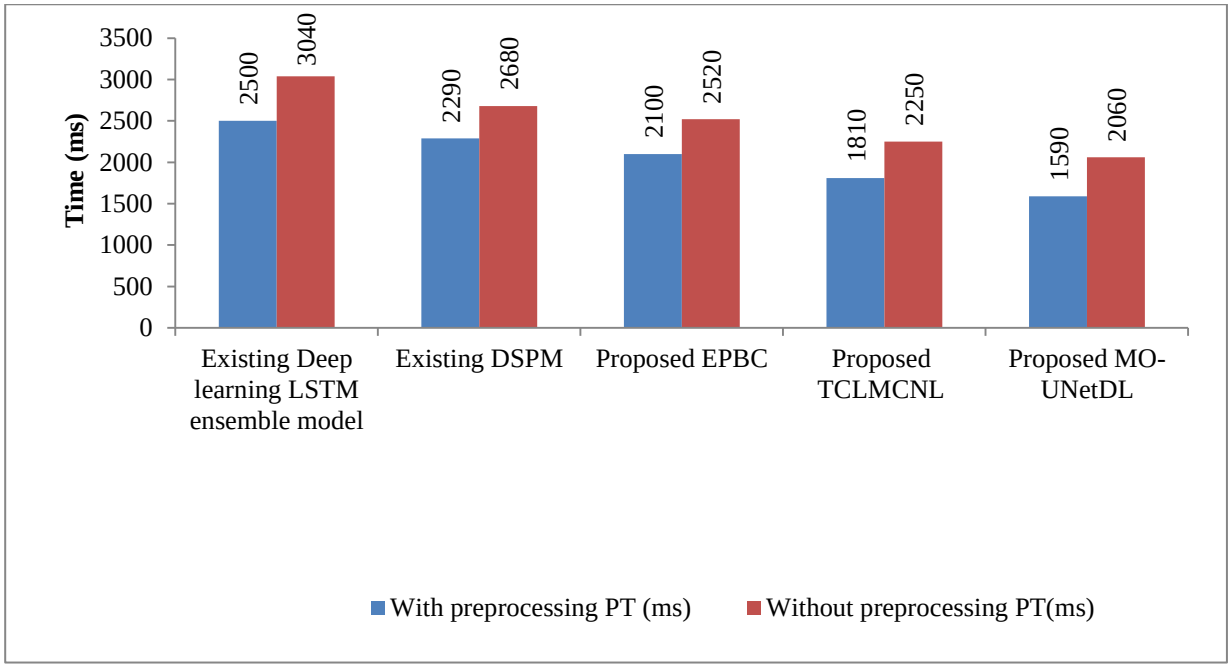

Figure 8.1 (f) Performance of Prediction time with and without Preprocessing model for Pneumonia dataset

Figure 8.1 (a) to (f) shows the classification performance of five different methods with and without preprocessing methods for COVID-19 and Pneumonia datasets. By observing the above graphical values, it is evident that accuracy, precision, recall, and specificity are significantly improved, and time is minimized in the proposed MO-UNetDL technique with the preprocessing method compared to other methods. The higher accuracy, precision, recall, and specificity, along with a lesser time, are achieved by using a MO-UNetDL classifier. The classifier uses the selected features from the preprocessed data as input. With the input, accurate classification is achieved by identifying the association between two datasets (i.e., training and testing) to classify the data accurately using Wilcox's index coefficient. The use of Wilcox's index coefficient for measuring dataset association contributes to improved accuracy, precision, recall, and specificity, while minimizing time. Therefore, the accuracy of the proposed MO-UNetDL technique, combined with the preprocessing method, was evaluated using Equation 3.12, is improved by 9% and 6% compared to the existing deep learning LSTM ensemble method and DSPM, precision is increased by 8% and 6%, recall is increased by 9% and 6% and specificity is increased by 13% and 11% for the COIVID-19 dataset. In addition, the accuracy of

the proposed MO-UNetDL technique without the preprocessing method is increased by 9% and 8%, precision by 10% and 8%, Recall 10% and 8% and specificity is enhanced by 10% and 8% related to the existing LSTM and DSPM respectively for the COVID-19 dataset. Additionally, time for MO-UNet technique is decreased by 28% with preprocessing and its 25% without preprocessing for COVID-19 dataset when compare to the existing DSPM method.

The accuracy and precision of the proposed MO-UNetDL technique, with the preprocessing method, increased by 9% and 6%, the recall is increased by 8% and 6% and the output of specificity is enhanced by 13% and 10% when contrasted with existing methods LSTM and DSPM respectively for the Pneumonia dataset. Moreover, without preprocessing method the accuracy of the proposed Mo-UNetDL technique is 10% and 8%, the output of precision is 10% and 9%, the recall is 10% and 7% and additionally specificity is 10% and 7% when compared to the existing methods LSTM and DSPM respectively for the Pneumonia dataset.

8.3 CHAPTER SUMMARY

The performance of proposed preprocessing methods, such as the ALRTOHE method and ZMFNE method, feature selection methods, including the NSPPS method, TSIDFE method, and SCTPP-FS method, and classification methods, including the EPBC method, TCLMCNL method, and MO-UNetDL method, is evaluated using the COVID-19 and RSNA Pneumonia datasets. As discussed in the above results, the performance of preprocessing in the ZMFNE method is improved over the ALRTOHE method, achieving an accuracy of 88%, an error rate (ER) of 12%, a size of 48 KB, and a preprocessing time of 3830 ms for the COVID-19 dataset. Moreover, the outcome for Pneumonia dataset in accuracy is 90%, an error rate is 10%, a size is 46KB and time as 3630ms for the ZMFNE method.

The outcomes of feature selection methods are examined, and the SCTPP-FS method gives better performance when selecting relevant features from the preprocessed dataset than the NSPPS method and TSIDFE methods. The results of two sets, namely feature selection with and without preprocessing methods, are evaluated. The compared results show that the results of the SCTPP-FS method, with an accuracy of 91%, an error rate of 9%, a space complexity of 55 KB, and a feature selection time of 4950 ms for the COVID-19 dataset and for Pneumonia the

accuracy is 93%, an error rate of 7%, a space complexity of 50KB and the outcome of time is 4900ms. The experimental results of the classification EPBC method, TCLMCNL method, and MO-UNetDL method using with and without preprocessing methods are evaluated. The obtained results shows that the better classification results are obtained in third proposed MO-UNetDL method. The following chapter shows the results and discussion of Machine Learning and Deep Learning techniques using both 70:30 and 80:20 data splits.