CHAPTER 10

CONCLUSION AND FUTURE DIRECTIONS

10.1 Conclusion

This research developed novel machine learning (ML) and deep learning (DL) methods for disease prediction. The main contributions involved in this work are divided into three parts: preprocessing, feature selection, and classification. To get standardized information for ease of use, data preprocessing methods are incorporated that perform well to provide consistent data. The dimensionality issue is addressed in this research by designing novel feature selection methods. Additionally, data classification methods have been developed to categorize data with reduced error, thereby improving the accuracy of disease forecasting.

The first part of the work proposes performing data preprocessing using two new preprocessing techniques: the ALRTOHE technique and the ZMFNE technique. First, the proposed ALRTOHE technique is designed to eliminate noise in the dataset accurately through additive log-ratio transformation and the one-hot encoding method. By applying the additive log-ratio transformation, input data is standardized. This helps ensure that each dataset has a similar format, providing consistency for further processing. One-hot encoding is employed in the ALRTOHE technique to transform numerical categorical variables into binary vectors. With this, the binary representation of data is obtained. Thus, the proposed ALRTOHE technique for COVID-19 and Pneumonia datasets achieves an average preprocessing accuracy of 13% and 15%, a lesser preprocessing time of 11%, a space complexity of 12% and 11%, and an error rate of 35% compared to the existing methods, was evaluated using Equation (3.12).

Another contribution to preprocessing input data, known as the ZMFNE technique, is developed for removing noisy data and addressing dimensionality issues. The data is initially converted into a matrix, and zero-mean feature scaling is applied to normalize the information. The data transformation process is performed using one-hot encoding to preprocess the information. With this pre-processed data, time and space complexity is reduced. The outcomes of the ZMFNE method are compared with existing methods. The obtained output demonstrates that the preprocessing accuracy of the ZMFNE technique for COVID-19 and Pneumonia datasets achieves an average preprocessing accuracy of 17% and 18%, with minimum time of 17%, a space complexity of 18% and 14%, and an error rate of 48%, 45% associated to the current approaches.

The second part involves a feature selection process, proposing the NSPPS, TSIDFE, and SCTPP-FS techniques. The proposed NSPPS technique is employed to select error-minimized, optimal, and relevant patterns, thereby improving accuracy. Data in an elevated-dimensional space is charted to a minimum-dimensional space using Sammon projection for managing the inter-point distance structure. Nonlinear Sammon Projection is used to determine the pertinent and redundant patterns. The error-reduced, pertinent patterns are only considered for disease prediction with less time. The performance of the NSPPS technique achieved an average Feature selection accuracy of 5% and 4%, a reduction in time by 13% and 31%, a reduction in space complexity by 10% and 12%, and increase in ER by 22% compared to conventional methods.

The TSIDFE Technique is designed to select features while reducing the error rate. The computation of the Tversky index similarity between features in the dataset aids in recognizing both pertinent and irrelevant features. Accurate identification is made by considering the similarity coefficient where ‘0’ specifies the feature is irrelevant and ‘1’ specifies the feature is relevant. Relevant features are only taken for further processing. This reduces the feature selection time by 20% and 43%, space complexity by 16% and 22%, and error rate by 37% and 38% and feature selection accuracy by 8% and 6% higher compared to conventional works for COVID-19 and Pneumonia datasets.

The SCTPP-FS method is designed to choose important features for disease forecasting. The Kaiser–Meyer–Olkin correlation is used to select the target or essential features. Correlation among features is calculated and determines relevant and irrelevant features. Through correlation analysis, the most significant features for accurate disease forecasting are identified. The results of feature selection accuracy are improved by 11% and10%, with 26% and 47% reduction in feature selection time, a 24% and 28% reduction in space complexity, and a 48% and 57% reduction in error rate compared to existing techniques for COVID-19 and Pneumonia datasets.

The third part of the research work executes the classification process by proposing the EPBC technique, TCLMCNL technique, and MO-UNetDL technique for disease prediction. The proposed EPBC utilizes a perceptron binary classifier to classify the input data through a weighted sum. Emphasis boosting is created through the weighted sum that splits the patterns or data with zero training error. With this, accurate prediction is achieved, with an improvement of better precision by 2% and 3%, recall by 1%, and specificity by 8%, and reduction in prediction time by 16% and 15% compared to conventional methods for COVID-19 and Pneumonia datasets.

The proposed TCLMCNL Technique is developed to perform data classification and predict disease with reduced error. Different types of layers are used in TCLMCNL, where each performs a distinct process. The time-dependent Cox regression calculates the relationship among data using Carmer's phi correlation function. Also, Huber loss is estimated to provide minimum error classification results. The results of prediction accuracy by 2%, precision by 3%, recall by 2%, and specificity using the TCLMCNL Technique are increased by 12% and 11% and reduction in prediction time by 25% and 31% compared to the existing methods for COVID-19 and Pneumonia datasets.

Proposed MO-UNetDL method is designed to improve result of disease prediction through classification. Wilcox's index coefficient is employed to discover similarity among input data. Also, max-pooling operation is performed to minimize the data dimension. Later, data classification results are acquired. By using memetic optimization, hyper-parameters are tuned to get optimized classification results through executing truncation selection, Two-point crossover and through bit flip mutation. Therefore, the outputs of prediction accuracy increased by 3%, precision by 4% and 3%, recall by 3%, specificity using MO-UNetDL Technique are improved by 21% and 20% with minimum prediction time by 27% and 48% than conventional methods for COVID-19 and Pneumonia datasets.

10.2 Limitation and Future Scope

As seen in the proposed techniques, numerous efforts have been made to identify diseases at an early stage. However, it is still a tough job to identify fewer, more specific features that can predict disease extremely early. Future work will focus on extraction as well as the study of novel features that are more likely to assist in the recognition of disease. Additionally, features found to be less informative within the proposed framework will be excluded to enhance prediction performance. This would help doctors, physicians, and healthcare professionals make effective decisions for disease diagnosis.

The proposed methods involved applying an optimization technique to adjust the fine-tuning hyperparameters. But, it was unable to select optimal features with lesser error. Future work will extend the proposed techniques by incorporating novel optimization techniques into deep learning (DL) methods for categorizing a variety of data with minimal prediction error by adjusting the hyperparameters. This will enable the proactive recognition of diseases and significantly improve efficiency in the healthcare industry by facilitating faster and more accurate diagnosis.

The proposed techniques were not considered, and the computational cost parameters for disease detection were not estimated. Additionally, the computational cost associated with handling a large number of data samples will be analyzed in the future.

The proposed techniques did not sufficiently address the automatic prediction of patient data, considering higher security. Additionally, the automatic prediction of patient data requires increased security, as patient medical information is highly confidential. Therefore, in the future, security models can be used to build prediction models for securely diagnosing patient data.

The proposed methods lacked consideration of other medical domains, namely pattern recognition and object detection, for precise image classification. This contributing to the improvement of medical diagnostics and patient care. In future work, DL-based pattern recognition and object detection will be utilized. Pattern recognition will be utilized for identifying recurring patterns and features within images. Object recognition will be utilized to find specific objects within an image, such as organs, tumors, or blood vessels. This can be helpful in early disease detection.