CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

The disease prediction method considers a predictive model to predict symptoms based on the disease. Prediction models assist healthcare consultants and patients in making medical decisions. In order to encourage clinical decision-making with better patient outcomes and quality of care, the prediction method results in precise patient risk tier. Also, upcoming health-related occurrences are determined through medical risk prediction models. DL-basis CNN models, as well as Machine Learning algorithms, are applied to disease prediction. Several existing disease prediction methods perform well, but accurate results are not obtained. Many research works were introduced for early disease prediction. Likewise, many have been introduced to develop optimized machine learning and deep learning models for predicting disease with huge prediction accuracy in less time.

2.2 LEARNING METHOD BASIS OF DISEASE PREDICTION

A Novel Feature Reduction (NFR) model incorporating Machine Learning (ML) as well as Data Mining (DM) algorithms was proposed by Syed Javeed Pasha *et al.* (2020) to reduce the error rate and stimulate the performance rate. In order to attain potent and powerful disease prediction, the NRF technique employed two dissimilar methods. The first method was based on the heuristic process by decreasing the feature requirement for further processing. Another method is obtaining all the specific features to ensure accuracy, thus decreasing functioning time to a greater extent. The designed model was not applied to an extensive range of disease datasets.

A smart health monitoring for heart disease prediction via ensemble deep learning and feature combination to increase the accuracy involved during predicting was proposed by Farman Ali *et al.*,(2020). The sensor data extracted features were fused through the feature fusion method, and electronic medical records produced valuable healthcare data. Within the data, abnormality features were eliminated through the information gain technique, and significant features were picked to increase the system's performance and lessen the

computational burden. The meta-learning classifier was employed in this technique to attain high accuracy in classification. The missing values were not efficiently handled.

A machine learning approach was employed by Sanjay Kumar Sonbhadra *et al.,*(2020) for deriving the activities and trends of disease-relevant problems which a patient is affected with. A holistic approach to the epidemiological features and information of infection was critical to control its spreading. The proposed method with reduced features generates significant outcomes that are verified through the cosine similarity score. For analyzing and detecting disease, a review of ML and DL methods was investigated by T. Aishwarya and V. Ravi Kumar (2021). The CheXNet architecture-based detection accuracy was highly attained, along with precise predictions. For differentiating kinds of diseases, computerized investigation and analysis were developed. It failed to develop advanced disease detection because other prediction methods were not used.

ML was proposed for outbreak of illness prediction by Aman Khakharia *et al.,*(2020) for two extremely dissimilar dense and population countries. An available optimized resource was producing high-accurate prediction. It also helped to reduce the cost of managing the pandemic in addition to enhancing the recovery process inside regions. The role of ML algorithms employing ML techniques was designed by Ameer Sardar Kwekha-Rashid *et al.,*(2021) for augmenting testing accuracy involved in prediction. In addition, it evaluates the cases of the disease presented. Different ML models were employed to offer better prediction outcome. The error rate was not considered.

The descriptive survey approach was conducted by Kailash Nagar *et al*, (2022) for assessing prevalence of communicable and non-communicable diseases and evaluating standard living among rural populations through face-to-face data collection in particular area. The main drawback is its reliance of self-reported data.

An automated COVID-19 diagnosis and classification using CT portraits, was developed by Dac-Nhuong Le *et al*, (2021) for integrating IoT for real-time data collection and monitoring, but it is not feasible in all health care environments.

A study on Digitization of healthcare sector was developed by Metty Paul *et al*, (2023) to analyze the impact of digital transformation in healthcare to focus on identifying privacy and security challenges through electronic systems but it is vulnerable to susceptible fraud and illegal actions.

A framework used to automatically detect and classify lung diseases was developed by Shimpy Goyal *et al,* (2021) specifically for COVID-19 and Pneumonia to address the challenges of viral and bacterial causes using multi-stage feature extraction and classification approach, which provides high accuracy but it requires high computational resources.

The incorporation of three DL models was developed by H. Abbasimehr and Paki R. (2021) for predicting disease. The multiple-output forecasting design method was predicting multiple time points. For each model, the Bayesian optimization method picks the finest hyperparameters and increases prediction performance. However, the designed method was ineffective during useful feature selection over time series and in incorporating them in DL strategies.

Many applications of DL schemes were introduced by Shorten C *et al.*,(2021) for disease prediction. The performance ineffectiveness was discussed. Additionally, it offers a detailed analysis of input data. Kallel A et al. (2022) introduced a federated machine learning model for disease prediction. This model ensures consistent decision-making through federated batch ML. Also, it learns and trains the gathered information. It failed to augment accurate disease prediction by means of DL methods.

Internet of Things (IoT)-basis DL method was developed by Ahmed I *et al.*,(2021) to predict disease. It effectively minimized medical experts/radiologists' workload and contributed to pandemic control. During disease forecasting, the Region Proposal Network (RPN) was utilized. However, it struggled to handle the large amount of data needed for accurate predictions. Chandra R *et al.*,(2022) introduced Recurrent Neural Network models for predicting multi-step disease infections. It uses LSTM models for forecasting the spread of infections, and it also includes univariate and multivariate time series forecasting methods. The designed

approach-based prediction was designed to offer robust predictions, but accurate forecasting was not performed in less time.

A deep-LSTM ensemble model was designed by Shastri S *et al.*,(2021) to diagnose disease. In addition, it augments medical healthcare facilities with AI intervention. Besides, AI-aided automated healthcare systems were used to treat infection contactless. Also, it offers AI-based self-treatment in remote locations. However, the ensemble method did not reduce the error rate. Recurrent CNN schemes were introduced by Sakthivel R, *et al.*,(2022) to capture the multifaceted improvement of disease occurrences to attain disease prediction. The dissimilar kinds of deep learning models were employed for fine-tuning along with improving disease detection performance. However, it did not improve the prediction accuracy.

An ensemble deep learning approach was created by Abedin MZ *et al.*,(2021) for fusing Bagging Ridge (BR) by means of bi-direction along short-term memory (Bi-LSTM) neural networks with base regressors with the Bi-LSTM BR method. This approach aimed to improve prediction performance and identify the minimized rick impact during the pandemic However, it failed to decrease the error rate.

ML methods were introduced by Solayman S *et al.*,(2023) for examining automatic disease detection. For the automatic prediction model training, Preliminary analysis and data preprocessing were executed. Hyperparameter tuning was performed to discover the salient hyperparameters. A hybrid DL model was implemented to tackle the progression-based symptom features of the employed disease dataset. It failed to augment Prediction accuracy.

DL framework was developed by Cui Z *et al.*,(2023) for electricity demand prediction. A multiple linear regression model was implemented for the disease-related variables. An efficient DL algorithm was used to remove diseased data impacts. However, the designed framework did not accurately detect the disease prediction. DL algorithm was used by Mareeswari V *et al.*,(2022) to detect the disease. The approach intended to learn about the automated diagnosis systems process and to learn about AI's basic methods used in this process. It is observed that a processing-based diagnostic system generates improved outcomes with respect to illness. Yet, it didn’t succeed in the reducing prediction time.

A new deep transfer learning techniques named “COV-DLS”, introduced by Kumar V *et al.*,(2022) to improve accuracy. The overfitting issue was to overcome through the dropout layer. In order to attain high-performance efficiency, Modified DL architectures were designed. The designed model was designed to attain high accuracy as well as an F1 score. It failed to execute preprocessing. DL method was developed by Kafieh R *et al.*,(2021) for precise prediction. It failed to lessen the severity of patient outcomes. DL models were analyzed by Aftab M *et al.*,(2022) to discover normal, influenza, and other diseased cases. In order to find whether the input was normal or abnormal, efficient classifiers were developed. However, it did not detect the relevant feature accurately.

A hybridization of Graph Convolutional Network as well as Gated Recurrent Unit models was proposed by Amgad Muneer *et al.*,(2022) for the mRNA degradation that predicts the mRNA sequences degradation risk. Also, it effectively forecasts stability/reactivity. Moreover, it highlights the model's efficiency and efficacy. The designed model validation loss was not lowered. DL techniques were discussed by Xu L *et al.*,(2022) to forecast disease. During performance prediction, significant development augmented by it. The combination of DL methods was to help effectively handle disease prediction. Moreover, calculating the spread of the infection, the effectual methods were developed. It did not provide a better prediction outcome.

A deep CNN architecture was proposed by Aijaz Ahmad Reshi *et al.*,(2021) for the chest X-ray image classification basis disease diagnosis. To augment the CNN model performance, Data augmentation techniques were developed. During image classification, data augmentation methods provided CNNs with the ability to remain unchanged. This approach worked well during pandemics. The accuracy of disease predictions improved, and the loss in the learning model decreased. However, it did not take into account the performance time.

The supervised machine learning model was introduced by M.T. Huyut, (2023) to improve existing accuracy by using patient data, the infected patient groups were detected. It helped sort out that the feature dataset comprises of Routine-Blood-Values (RBV) and demographic data that affect the disease prognosis. Time complexity was not decreased. Parul Arora *et al.*,(2020) introduced deep learning models to improve accuracy. The designed

model was implemented to predict positive disease cases with fewer errors. In order to minimize the spread in particular zones, preventive computing was invented. The relevant attributes were not selected.

The automatic detection and diagnosis of COVID-19 from chest CT scan images using 3D-CNN was developed by Mahboob *et al.,* (2021) to improve diagnostic accuracy, which achieves better classification accuracy than traditional 2D-CNN, but the method requires high memory and processing power.

The review and analysis of the potential applications of artificial intelligence in healthcare by Lina Sun *et al.,* (2021) developed to review the current applications and future potential, provides a broad survey of AI applications but not address the domain–specific details.

The automated prediction of heart disease using an improved explainable learning-based technique by Zuping Zhang *et al.,*(2024) was designed to enhance the accuracy and it outperforms existing methods, but it was based on single dataset and lack of clinical deployment.

However, another DL approach was presented by Alassafi MO *et al.,*(2022) that combined RNN as well as LSTM networks. The designed approach was used to enhance prediction accuracy with less error. Also, it predicts the time series well, but it does not organize the database. The faster R-CNN and mask R-CNN methods were presented by M. Emin Sahin *et al. (2022)* to train and test the database for classifying infection types. With this intention of the method was able to precisely diagnosing disease by means of CT images. The performance time was not computed.

Disease prediction with the assistance of the DL technique named LSTM by Zhifang Liao *et al.,*(2021) can also be seen. The hybrid infectious diseases prediction model helped address prediction challenges. It was used in real-time the forecast the trend of infectious diseases during the current large-scale vaccination efforts. It failed to reduce the disease prediction error.

LSTM Variational Autoencoder was developed by Farhan Fuad Abir*et al.,* (2022) to identify infection. In this approach, wearable device dataset was trained and assessed. Also, it effectively detected abnormality among the infected individuals during a pandemic. However, it

failed to analyze performance time. The disease prediction was handled by the AI technique in Swati Swayamsiddha *et al.*,(2021). It was an automatic disease detection system that tracked and treated patients remotely. However, the smart-health care industry was not implemented to offer rapid and more accurate outcomes.

CNN-GRU- based hybrid learning model created by Anil Utku (2023) aims to predict the spread of disease. In this model, CNN conducted convolution and pooling operations to pull out spatial features from the input data. Also, GRU provides long-term as well as non-linear associations for learning. It failed to increase the disease prediction accuracy. VOC-DL prediction framework developed by Zhifang Liao *et al.*,(2022) to predict the daily new confirmed cases. The time series dataset includes VOC variant data, and it implements the slope feature method in the designed prediction method. However, the other variants' prediction was not performed by the designed framework.

A new ML forecasting method was suggested by R. Sujath *et al.*,(2020) to predict disease transmission. Their goal was to identify the epidemiological occurrence of the illness and the rate of disease cases in India. They used liner regression, multilayer perceptron, and vector autoregression model on disease data. It did not consider the prediction time. Another ML model was introduced by Furqan Rustam *et al.*,(2020) to predict the number of patients affected by the disease. They implemented the forecasting techniques, including linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES), to predict the disease factors. ES performed best in forecasting new confirmed cases, death rates, and recovery dates. However, the existing forecasting methods had higher computational costs.

A machine learning approach to symptoms basis disease diagnosis was proposed by Yazeed Zoabi *et al.*,(2021). During the full impact of the disease, the designed model was trained on personal data in tested conditions for SARS-CoV-2. Also, it applies globally for effectual screening and testing prioritization for the infection inside the public. It was unable to improve the overall work accuracy. A review of artificial intelligence was investigated by Chellammal Surianarayanan *et al.*,(2021) to evaluate individuals' risk levels during disease detection. In

addition, it portrays a variety of challenges that are linked with the execution of the AI-based tool. However, the performance time was not diminished.

Ahuja S *et al.,*(2022) introduced a deep learning method that used Convolutional Neural Network (CNN) and a stacked bidirectional fated recurrent unit (Bi-GRU) to recoganize and detect disease. Also, the designed model was augmented towards a positive case recovery rate. The designed process increased prediction accuracy, but the performance time was not reduced. A deep-LSTM ensemble model was developed in Ketu S, Mishra PK (2022) to forecast illness with better accuracy. The significant data was extracted using convolutional layers and enriched with the ability of the LSTM layer. Deep-leaning-based solutions are also offered for economic and social prosperity. However, it does not show the positive results achieved in a short period.

A novel LSTM deep learning building was designed by Lucas B *et al.,* (2022) for disease prediction. It offers more accurate predictions. The spatial association strength was described through Facebook's Social Connectedness and Movement Range datasets. However, the designed architectures did not deeply analyse the features for precise prediction. ML and DL-based disease pandemic method was introduced in Showmick Guha Paul *et al.,*(2023) to develop understanding in addition to manage the illness situation. The designed model enhanced the disease prediction accuracy, although it is not suitable for diverse data formats of disease datasets.

Machine learning approach was developed by K. Arumugam *et al.,*(2023) to predict several kinds of diseases. Data visualization with the ML method determines Diabetes-related heart disease that arises due to diabetics. During the prediction of heart disease within diabetic patients, the decision tree model was used to attain optimum performance. An analysis of virulent disease detection as well as prediction carried out by Sunday Adeola Ajagbe & Matthew O can be seen. Adigun (2024) using Deep learning methods. With the aim of the approach, minimal complication was aimed at attaining accurate as well as efficient performance with minimal complication. A comparison of existing methods is shown in Table 2.1.

Table 2.1 Comparative table of existing Learning Methods for Disease Prediction

Author name, Year & Ref. No	Method	Contribution	Datasets	Merits	Demerits
Farman Ali *et al.*,(2020)	Smart health monitoring for heart disease prediction	Smart health monitoring was developed for heart disease prediction	Heart disease Datasets namely Cleveland and Hungarian	Accuracy was improved	Missing values were not handled
Kumar V *et al.*,(2022)	Deep transfer learning techniques	New deep transfer learning techniques named "COV-DLS" were examined with higher accuracy	Covid -19 dataset	Higher performance efficiency	It failed to perform preprocessing
Sakthivel R ,*et al.*,(2022)	CNN schemes	CNN schemes were developed for disease prediction	COVID-19 Radiography Database	Overhead was minimized	CNN failed to enhance prediction accuracy
Sanjay Kumar Sonbhadra*et al.*,(2020)	Machine learning approach	Machine learning approach was utilized for extracting the features	COVID-19 dataset	Time was lower	Accuracy was not improved
Shastri S *et al.*,(2021)	Deep-LSTM ensemble model	Deep-LSTM ensemble model was used for diagnose disease	COVID-19 datasets	Recall was increased	Ensemble method did not reduces error rate
Syed Javeed Pasha *et al.*, (2020)	NFR model	NRF method employed for attaining the efficient disease risk prediction	Cleveland and Hungarian heart dataset	Error rate minimized	NFR was not employed on extensive range of disease dataset
T. Aishwarya & V. Ravi Kumar (2021)	ML and DL methods	Review of ML and DL methods were examined for disease	COVID dataset	Detection accuracy was higher	It failed to develop advance disease detection

2.3 DIESEASE PREDICTION FEATURE SELECTION MODEL

A step-by-step model using sterilization, feature selection and classification was proposed by Srinivas Koppu *et al.*,(2020). In order to optimally tune multiplied weights and decrease the error, a modified Dragonfly Algorithm was employed. However, the proposed model is on high-quality datasets, but efficiency and reliability were not ensured. A new way to represent hybrid features was introduced by ChunyanAo *et al.*,(2020) for forecasting the antioxidant protein level. The designed model incorporated hybridized random forest and hybrid features. In order to carry out classification, feature extraction and hybrid feature representation were utilized. The feature selection was not executed in a precise fashion.

A hybrid deep learning model was developed by Mohan Abdullah *et al.*, (2024) with maximum accuracy and lesser feature dimension. It was utilized to help radiologists and physicians for handling misdiagnosis rates.

A case-based reasoning framework was proposed by Olaide N. Oyelade & Absalom E. Ezugwu (2020) for early disease detection and diagnosis by utilizing semantic-depended mathematical modeling for minimizing the diagnosed case false positive rate. An ontology learning algorithm carried out the patient data feature extraction and mapping. Furthermore, it effectively detects or classifies diseases, whether they are present or not. However, the performance of different similarity calculations was not performed.

A novel SMbGI framework was developed by V. Laxmi Narasamma *et al.* (2022) to discover malware activities. In the beginning, it collected the "disease vaccine tweets" from Twitter, and the irrelevant tweets were removed during preprocessing. Then, feature extraction was performed to eliminate the aspect terms. Moreover, the designed framework effectively identified malicious activities. Lastly it carries out sentiment classification on the tweets. It failed to increase the prediction accuracy. A comparison of existing methods is shown in Table 2.2.

Table 2.2 Comparative table of existing Feature Selection Models for Disease Prediction

Author name, Year & Ref. No	Method	Contribution	Datasets	Merits	Demerits
ChunyanAo *et al.*,(2020)	Hybrid feature representation method	Hybrid feature representation method was designed to predict the antioxidant protein level	Non-antioxidant protein database	Accuracy was higher	Feature selection was not executed
Linghua Wu, *et al.*,(2024)	Deep Learning based anchor-free detector	Deep Learning based anchor-free object detection framework	Pneumonia RSNA dataset	Error rate reduced	Failed to work in large dataset
Olaide N. Oyelade& Absalom E. Ezugwu (2020)	Case-based reasoning framework	Case-based reasoning framework was executed for early disease detection	COVID-19 dataset	False positive rate was reduced	Performance of different similarity calculation was not performed
Srinivas Koppu *etal.*,(2020)	Step wise model	Step wise model was utilized for cleaning, feature extraction and classification	Cleveland, statlog and wisconsin database	Accuracy was enhanced	Reliability was not ensured
V. Laxmi Narasamma *et al.*,(2022)	SMbGI framework	SMbGI framework was developed to find the malware behavior	Twitter dataset	Time was lower	Failed to enhance the prediction accuracy

2.4 DISEASE CLASSIFICATION ALGORITHMS

Cardiac disease prediction aided the doctors in creating exact conclusion regarding the patient health. A dimensionality cut down method was proposed for detecting the similar heart disease features by means of feature selection model by Anna Karen Garate-Escamila *et al.*,(2020). It failed to present prediction accuracy and space complexity as they were not diminished. Disruptive technologies for analyzing disease using several decision-making models were tested by Mohamed Abdel-Basset et *al.*,(2020) to perform effective diagnosis. Also, it restricts the spread of infection and ensures healthcare security. Furthermore, it adopts methods to lessen the unprecedented outbreak's impact of the disease.

To predict the rise of the disease in Egypt by Mohamed Marzouk *et al.*,(2021) used artificaial intelligence. This disease was highly contagious and showed both symptomatic and asymptomatic patterns in infected patients. As, a result, the number of active cases increased quickly. This also sped up the process of detecting and diagnosing the disease worldwide. However, the study did not account for the effects of non-pharmaceutical measures on the spread of disease.

However, another method for combining features, which fuses matrix-based representation with Convolutional Neural Networks (CNN), was proposed by Haolin Wang *et al.*,(2020). This method aims to extract features and combine them for effectively using multiple data sources in medical-decision making. The goal was to achieve high accuracy in clinical data mining, but the reliability of the performance did not improve.

A study looked into immune-related relationships in specific diseases proposed by NopharGeifman& Anthony D. Whetton (2020). This designed approach scrutinizes immune-related molecular and cellular patterns within the context of the pandemic. The immune system's underlying disease-related complications were not considered. A method for predicting disease time series worldwide was proposed by Patricia Melin *et al.*,(2020). It uses a hybrid ensemble modular neural network that includes non-linear autoregressive neural networks. This approach removes different and unnecessary feature, which helps reduce the burden of making predictions. A fuzzy logic system was not applied in the designed prediction method.

An innovative IoT as well as Cloud basis Blockchain Model was introduced by Ahmed S. Salama *et al.*,(2022) for disease detection. Moreover, it controls and lessens the spread of infection. In addition, it effectively handles the smart contract rules based on different personal conditions in the blockchain system. However, the performance of the prediction error rate was not reduced.

To precisely find the abnormal activity prediction, Deep Learning depend LSTM Models were designed by Manju D *et al.*,(2022). It employs a Recurrent-ResidualInceptionV3 network to decrease the implementation time. In order to achieve high efficiency, the performance load was minimized through the wider network and deeper inside the training phase. The error was not lessened.

A complete review of chronic disease prediction using machine learning algorithms was presented by Rakibul Islam *et al.*,(2024). It especially diagnoses individuals with chronic diseases. The error rate needs to be minimized.

3D Convolutional Neural Networks were introduced by JV Vardhan *et al.*,(2021) to attain earlier as well as exact outcomes of disease diagnosis. This system uses ResNet-50 to acquire predictions on the 3D CT images. Moreover, the designed network method was in contrast to other deep learning models and combination techniques. The designed network prediction time was not reduced. Granular network traffic classification technique was developed by Zaki F *et al.*,(2021) to generate classification at two granularity levels using classifier chain. In order to ensure better performance estimation, the prequential assessment method was used in the designed technique. Real-time traffic classifiers were not developed using this approach.

A multi-task Gaussian process (MTGP) regression approach was introduced by Ketu S & Mishra PK (2021) to accomplish better disease forecast. The method aimed to predict the disease outbreak globally. Also, it helped to offer preventive measures to diminish the impact of the spreading communicable disease. However, it failed to predict the MTGP with deep feature learning to improve the model's capacity.

A hybrid approach was developed by Ossa LFC *et al.*,(2021) for disease prediction, in this approach that fuses recurrent neural networks-based SIR model differential equations. It also

detects structural modifications inside the method. A depth-wise separable convolution neural network (DWS-CNN) was introduced by Le D *et al.*,(2021) using a deep support vector machine (DSVM) for disease detection. The designed method identifies both binary and multiple classes of disease. In data acquisition, devices collect patient data and send it to the could server. Then, the Gaussian filtering was applied to perform data preprocessing. After that, the relevant features were picked to categorise disease binary and multiple class labels via the DSVM model. However, it failed to attain maximum classification outcomes efficiently.

Iwendi C *et al.*,(2021) introduced ANFIS for early disease detection. It also employs uncertain systems for effectively forecasting global spread of the disease risk factor. Support Vector Machine classifier accurately classifies the disease dataset. The new variant of disease data was not applied in the system. Stacking an ensemble with deep neural network was created by Gupta A *et al.*,(2021) to predict post-disease symptoms. Then, long-lasting difficulty risk was predicted based of post-disease symptoms. For the early prediction of heart diseases, a stacking ensemble of deep neural networks was designed. This method lost in efficient feature selection by means of improved dimensional data.

The recurrent and CNN schemes were created by Hanuman Verma *et al.*,(2022) to predict disease-confirmed cases. Also, it deals with numerical modelling demands. The designed approach offers accurate prediction, but the error rate performance did not decrease. A Harris Hawk's optimization was progressed by Ye H *et al.*,(2019) to differentiate the disease severity—the method aimed to optimize the Fuzzy K-nearest neighbor. The designed model also outperforms further estimation-based competitors. However, the designed method failed to use other popular DL techniques to increase the disease prediction accuracy.

The logistic scheme was introduced by Bhardwaj R (2020) for predicting the evolution of the disease pandemic. A regression analysis was employed to predict the statistical model. In particular, the infection growth rate was fitted as an exponential decay. It failed to improve the accuracy. Adaptive neuro-fuzzy inference schemes, and improved beetle antennae search (BAS) algorithm, were implemented by Zivkovic M *et al.*,(2021) for disease prediction. The World Health Organization's official data was used in the designed prediction process. However, it did not reduce the prediction error rate.

LSTM networks basis of the transfer learning approach was introduced by Gautam Y (2022) to predict diseases. In this designed model single as well as multi-step prediction were carried out. Also, it is helpful for the identification of earlier infection spread prediction. However, it did not reduce the time needed for disease prediction. A novel cross-entropy basis loss function was employed by Anand Motwani *et al.*,(2023) to augment the CNN algorithm's convergence. Furthermore, it effectively evaluates huge patient datasets. Several nature-inspired optimization methods were implemented to increase designed model accuracy as well as DL algorithms during the earlier convergence. However, the risk factors still occurred.

Yujia Xu *et al.*,(2023) identified the augmentation techniques to understand the incremental levels and use the large open-access benchmark dataset. Additionally, the SR technique is derived from contrastive learning and applies CNNs. However, the designed detection offered inadequate accuracy outcomes. D. Ayris *et al.*,(2022) used a deep sequential prediction model (DSPM) and machine learning-based Non-parametric Regression Model (NRM)predict the spread of infection. The given disease dataset was trained and tested by DSPM- NRM. It assesses the designed model Mean Absolute Error. The model predicts the spread of infection well and has lower error rates.

Decision-making-basis federated learning network (DMFL_NET) was created by Malik et *al.*,(2022) in less time. The designed model was to collect information from various hospitals and then execute precise predictions with high security. The accuracy was not increased. The disease classification model was designed by Lin Wang *et al.*,(2022) using a fusion of patient demographic and comorbidity data. The feature-based time series data depends on the new Lasso Logistic Regression model to find out disease severity. The effective classification model was employed to enhance the effectiveness of patient treatment.

An automatic method was introduced by Hicham Moujahid *et al.*,(2022) for detecting and predicting patients' disease depending on their medical information. Within the classification outcome, the Gradient-weighted Class Activation Mapping technique was implemented for each class. In order to highlight the X-ray image regions of interest, the doctors easily interpreted the prediction outcome. The designed model's performance efficiency was also accurately evaluated.

MSTL method was developed by Sonakshi Garg *et al.*,(2022) for effectively forecasting infection. In addition, it relies on population density as well as economic conditions to manage the pandemic efficiently. Furthermore, the designed framework employs recurrent neural network architecture, LSTM a deep-learning model. This method did not augment improved disease prediction.

The medical score was acquired and reorganized by the patient's ML-based prediction in Carolin E. M. Jakob *et al.*,(2022) to identify patient's risk factors via a robust prediction method. The designed approach is based on disease prediction, which consumes much time. Dissimilar confirmed a diagnosis was detected, by means of hybrid model in Saqib Ali Nawaz *et al.*,(2021) depend on logistic models, when it applied on large amount of data it needed multiple adjustments. The method intended to explain the spreading dynamics of the disease. However, the prediction time was not into account.

The short-term prediction model was designed by Hongwei Zhao *et al.*,(2021) using daily incidence data. Poisson distribution was used in the observed incidence by gamma sharing for the series interval. During a short interval, effective reproduction values were computed. The untimely spoke about prediction of the disease was a critical problem in healthcare.

Hongbin Zhang *et al.*,(2022) introduced deep Contrastive Mutual earning (DCML) to identify the disease more effectively. The Fast Auto Augment-based multi-way data augmentation method aimed to improve the training dataset. This helps reduce the risk of overfitting. A new adaptive model fusion method was created to produce more distinctive image features. High-quality CT images were not accurately predicted. A Soft-Voting Ensemble Classifier was developed by Andrea Manconi *et al.*,(2022) for detecting the patients who suffered from a particular disease. The designed model was trained from 3D Inception-V1 and Inception-V3 CNNs. It failed to carry out data pre-processing to reduce the disease prediction complexity.

The AHEG-FS model was designed by Syed Javeed Pasha and E. Syed Mohamed (2022) to enhance disease risk prediction. An ensemble learning method was used to find important features. Then, the features were ranked via the gain ratio feature selection technique. Also, it

measures AUC with an accuracy of the feature reduction technique. According to attain most effective features subset, fewer contributing features were removed by means of the backward feature removal process.

ML, DL and transfer learning techniques were discussed by Hajar Lamouadene *et al.*, (2025) in diagnosing COVID-19. CNNs and Support Vector Machine (SVM) classifiers were employed to achieve higher accuracy in COVID-19 diagnosis. Transfer learning with ResNet18 combined to enhance the classification rates. However, the true positive was not enhanced.

The study of multimodal DL in medical diagnosis was focused on by Md Shofiqul Islam *et al*., (2025). Different studies of data sources, preprocessing, and challenges were discussed. Several DL techniques were employed in COVID-19 to explain methodology, data, and performance. DL with COVID-19 image, text, and speech data was examined for categorization. However, the time was not considered.

The CT COVID-19 model was implemented by Carlos Antunes *et al*., (2025) to help detect the disease early. DL models were introduced to accurately identify patterns in COVID-19 infections from CT-Scans. The designed model enhances the sensitivity and specificity. However, the prediction error was not minimized. Deep Learning- Based Segmentation for Chest X-ray Images was studied by Roshima Biju *et al*., (2025). They considered three CNN methods: U-Net, Resnet 18-UNet, and Mobilenet- Unet, to segment the lung regions in the chest X-ray images. In this way, the precision was improved. However, the ensemble methods failed to consider the diagnostic.

The binary classification approach was examined by Shirin Kordnoori *et al.,* (2025) with a convolutional structure. The feature was extracted with higher reliability. Various imaging conditions and patient demographics were illustrated. Also, the decision-support system was employed in a convolutional structure for offering accurate diagnosis and treatment for COVID-19. But, the accuracy did not improve.

However, another Machine Learning and Deep Learning methods like SVM and Feedforward Neural Networks, was used by Laura Verzellesi *et al.,* (2023) for COVID-19 mortality prediction. For evaluation, clinical features and only radiomic features are integrated

for classification. However, the prediction time was not reduced. A comparison of existing methods is shown in Table 2.3.

Table 2.3 Comparative table of existing Classification Algorithms for Disease Prediction

Author name, Year & Ref. No	Method	Contribution	Datasets	Merits	Demerits
Iwendi. C *et al.*,(2021)	ANFIS	ANFIS was introduced for early disease detection	COVID-19 dataset	Space complexity was minimized	Variant of disease data was not applied
JV Vardhan *et al.*,(2021)	3D Convolutional Neural Networks	3D Convolutional Neural Networks were designed to attain earlier disease diagnose	COVID-CT dataset and SARS-Cov-2 CT-Scan dataset	Precision was lower	Prediction time was not reduced
Ketu S & Mishra PK (2021)	MTGP regression	MTGP regression approach was examined for disease forecast	COVID-19 dataset	Feature time was lower	Failed to analyze deep feature learning
Dac-Nhuong Le *et al.*,(2021)	DWS-CNN	DWS-CNN was employed for disease detection	Chest X-ray (CXR) image dataset	F-measure was improved	Failed to efficiently attain maximum classification outcomes
Manju D *et al.*,(2022)	Deep Learning depend LSTM Models	Deep Learning depend LSTM Models were developed for precisely find the abnormal activity prediction	UCF-Crime dataset	Higher efficiency	Error was not lessened

Author name, Year & Ref. No	Method	Contribution	Datasets	Merits	Demerits
Mohamed Marzouk *et al.*,(2021)	Artificial intelligence based disease prediction	Artificial intelligence based disease prediction were used to predict disease in Egypt	COVID-19 dataset	Accuracy was improved	Failed to enable the impact of non-pharmaceutical intrusion of the spread of disease
Patricia Melin *et al.*,(2020)	Hybrid ensemble modular neural network	Hybrid ensemble modular neural network was employed for disease time series prediction	COVID-19 dataset	Error was reduced	Fuzzy logic system was not applied
ShimpyGoyal *el al.,* (2023)	Deep Learning Classifier - RNN with LSTM	Deep learning based model for accurately classifying COVID-19 and Pneumonia from chest X-ray images.	COVID -19 and Pneumonia dataset	Performance was high	Requires high computational resources for training

2.5 RESEARCH GAP

The hybrid deep learning model based on CNN-GRU is created to predict how the epidemic spreads. In the input data, spatial features were extracted via convolution, and pooling operations were used in the designed model. GRU provides long-term as well as non-linear association learning inferred by means of CNN. However, the reliability of disease prediction as the data volume increased. The confirmed cases of the patients were predicted using the VOC-DL prediction framework. VOC variant data included the time series dataset, along with the use

of slope feature method in the framework. Other variants that were not predicted by this system were dimensionality reduction, sensitivity, and specificity.

In order to perform improved disease risk prediction, the AHEG-FS model was designed. The ensemble learning feature selection technique was employed to find out significant features. In order to have prediction accuracy based on ranking features from higher to lower, a gain ratio feature selection technique was applied. Also, it measures the new feature reduction technique, AUC, accurately. The less contributing features were removed by the backward feature elimination method, thus attaining the most effective features subset. However, it is not suitable for various other disease dataset predictions.

The Chi^2-MI basis feature selection model was created to identify the high accuracy of Chronic Kidney Disease (CKD) and non-chronic kidney disease. In addition to this, the designed model has performed better implementation with the Extra Tress classifier. For CKD early patients, monitoring was executed via an ML-based designed model. However, time consumption in disease prediction has remained higher. Also, real-time kidney diagnosis with the assistance of a web application was not developed. To forecast disease, the LSTM-based deep learning method was introduced for measuring the proposed model efficiency as well as error rate. The disease precautionary measures were unavailable. DSPM was designed to forecast the disease spread with a minimum error rate. NRM was used to precisely and efficiently predict infection spreading range with poor prediction performance.

Several researchers have suggested various preprocessing methods to predict the trend of COVID-19. Also, the data normalization was not performed to reduce the space complexity. However, only using AI methods for prediction cannot capture the changing patterns of diseases over-time. A limited number of studies investigating hybrid models were combined with ensemble learning with deep learning techniques**.** The hybrid ensemble Feature Selection model using Machine Learning found in literature achieved effective feature subsets but had limitations in reducing the prediction time. Predicting the spread of the epidemic correctly is important for planning health management and developing economic and social action plan. Many studies in the literature examine and forecast the spread of COVID-19 in cities and countries. However, it failed to focus only on COVID-19-based variants and Pneumonia in the early stage and

minimised the error rate. To address the issue, novel proposed techniques are introduced to boost accuracy and diminish time and error.

2.6 THEORETICAL FRAMEWORK OF PROPOSED TECHNIQUES

The theoretical framework serves as the foundation for a research study's theory. These frameworks are essential to understanding important issues in healthcare systems. This study details the process of evaluating three theoretical frameworks for the early detection of diseases, specifically COVID-19 and Pneumonia as they emerge. The theoretical framework seeks to explain how to predict diseases during a pandemic to help safeguard human health. This work proposes a model for predicting COVID-19 and Pneumonia. The suggested model includes three stages: preprocessing, feature selection, and classification. The goal of this theoretical framework is to create a disease prediction system that uses ML and DL to enhance prediction accuracy.

The proposed system will address critical research gaps, including the accuracy of ML algorithms in predicting disease and the influence of environmental and lifestyle factors on disease risk. During the comprehensive literature review, data collection and preprocessing, feature selection and classification model development and experimental evaluation and validation, this study will provide the progress of an operative disease prediction system. The expected outcomes of this research include improved disease prediction accuracy, reduced error rate and enhanced clinical decision-making. This disease prediction system can help detect diseases early and prevent them, leading to better healthcare results for people and communities around the world.

2.7 CONTRIBUTIONS

To overcome the gap in the existing disease prediction models and feature extraction models, the main novelty and contribution of this research is as follows:

Initially, the Additive Log Ratio Transformed One Hot Encoding (ALRTOHE) and Zero Mean Feature Normalized Encoding (ZMFNE) technique is proposed for preprocessing. In order to attain high-accuracy in data preprocessing within minimal time, the ALRTOHE Technique is designed. The input datasets COVID-19 and RSNA Pneumonia are classified as disease datasets. Using this input, the ALROTHE model performs an additive log-ratio transformation and one-

hot encoding. IT employs a new method of log-ratio transformation to achieve data normalization. Following this, the model encodes the data to convert the numerical values into binary codes. The innovation of One Hot Encoding is to change numerical categorical variables into binary vectors. Binary representation of data is achieved with less time.

The proposed ZMFNE technique is developed to perform the input dataset preprocessing with high accuracy. This technique carries out the data normalization as well as the encoding process. The input data samples, as well as features, are taken from the COVID-19 and Pneumonia datasets. With this, a sample input matrix is generated in a row and column. Later, the novelty of zero mean feature scaling is implemented to regularize the input data. Subsequently, the data transformation process is done by one hot encoding. Here, the encoder alters the categorical features into a numeric array. The encoder considers the integer array-like or strings as input. Features are encoded, and a binary column for each class is generated. Accordingly, data normalization and transformation are processes performed to get the preprocessed output of data. Data normalization and transformation together performed to produce a high-quality preprocessed dataset which will form the effective foundation for accurate and efficient disease prediction.

Next, with the preprocessed data, 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 are proposed for feature selection. First, the NSPPS Model is developed to select the pertinent features with less error. The novelty of Sammon projection used in the developed model projects the features of high-dimensional space to low-dimensionality space. Then, the Patterns are chosen with Nonlinear Sammon Projection to find the disease as contagious during recently determined disease. From this, the inter-pattern distances are handled, thus diminishing the error rate while choosing the pertinent features. Thus, the NSPPS model enhances the feature selection performance. However, it remains a challenging task to identify a smaller number of more specific features that can accurately predict disease extremely early.

Another feature selection model termed TSIDFE is designed for feature selection to predict the disease. The preprocessed data is taken as input inside TSIDFE. With this input, the

Tversky index similarity coefficient among the two features is found. From this coefficient measure, the relevant and irrelevant features are detected. The Tversky index similarity coefficient between two features is estimated in the proposed TSIDFE technique. The similarity coefficient outcomes are varied from 0 to 1. If the similarity coefficient offers an outcome as '1', then the feature is pertinent. If the similarity coefficient gives an outcome of '0', then the feature is not relevant. Relevant features are selected for disease prediction. Irrelevant aspects are eliminated in the feature selection process. This improves accuracy and reduces time and space complexity. However, the accuracy of feature selection may be sensitive to the choice of similarity threshold in the Tversky index.

The SCTPP-FS technique was developed to recognize and pick pertinent features with better accuracy. For executing the selection process, the SCTPP-FS technique considers the dissimilar features as input. Then, the target features are mapped by calculating the correlation between the features. In the SCTPP-FS technique, correlation amid the features is computed with the assistance of Kaiser–Meyer–Olkin correlative projection pursuit for choosing the main features for accurate disease prediction. From the correlation computation, relevant and irrelevant features are found in less time. The relevant features are chosen with high accuracy and a lower error rate.

Finally, through considering the selected features as input, classification is carried out using the Emphasis Perceptron Boosting Classification (EPBC) technique, Time-dependent Cox regressive Levenberg–Marquardt Convolutional Neural Learning (TCLMCNL) Technique and Memetic Optimized U-Net Deep Learning (MO-UNetDL) classifier technique. The novelty of boosting algorithm is employed to obtain an accurate classification. The proposed EPBC technique employs the patient data with picked features as input. With this, the perceptron binary classifiers are formed by means of a weighted sum. The classifier divides the features and their data with zero training error. The weight and feature vector objective function is assessed to provide a predicted outcome. As a result, the actual classification outcomes are obtained using an accurate approach. However, achieving zero training error may increase the risk of overfitting in complex or noisy datasets.

The TCLMCNL technique was developed to categorize patient data to predict the disease. It comprises an input layer, an output layer, and hidden layers. Number of applicable features are taken as input in the input layer. The input is passed to the hidden layer. The TCLMCN uses time dependent Cox regression. It calculates Carmer's phi correlation function in the hidden layer. This allows for regression results and accurate classification. The Huber loss is also measured for both predicted and actual outcomes. In addition, the Levenberg-Marquardt algorithm has been used to reduce the error. Lastly, the classification result is obtained in the output layer with minimum error and high accuracy. The TCLMCNL technique lacks advanced mechanisms such as dynamic feature similarity evaluation and adaptive hyperparameter optimization, which may limit its flexibility and fine-tuning capability in highly complex disease datasets.

For disease prediction, the MO-UNetDL classifier model is proposed to classify the data with several layers. The input layer considers selected features from the database as input. MO-UNetDL technique determines the novelty of Wilcox's index coefficient to discover the similarity. Then, the index coefficient is offered to the innovation of soft step activation to calculate the index coefficient value in addition to providing the outcomes of either '1 or '0'. When the outcome of the activation function provides '1', the disease is correctly diagnosed. Besides, the innovation of max-pooling operation is used to minimize the data samples dimension by implementing the activation function. Upsampling is implemented to develop the dimension of data and provide classification results. Memetic optimization is employed to reduce the loss of data classification in MO-UNetDL for tuning hyperparameters. At first, the numbers of individuals are initialized. Fitness is estimated. The truncation selection, two-point crossover and bit flip mutation are measured. Once genetic operators are performed, the selected individual is measured, and fitness is confirmed. This process is repeated until an optimal solution is achieved. With this, the classification error rate is diminished and thus augments the accuracy.

2.8 CHAPTER SUMMARY

This chapter discusses the review of literature based on the existing disease prediction models, feature extraction models and contributions made with respect to disease prediction. The

research gap and contributions made on the same are also addressed to provide a comprehensive view of the need for and purpose of this research. The following chapter discusses the research methodology applied for precise disease prediction.