CHAPTER 1

INTRODUCTION

1.1 PRELIMINARY

Diseases are usually characterized by the signs and symptoms of a medical condition in a normal body. It is an irregular condition that affects the function of organisms. Both internal and external factors can cause these. For instance, internal immune system dysfunctions create various dissimilar diseases leading to multiple forms of illness, such as immunodeficiency, hypersensitivity, allergies, as well as autoimmune disorders.

Among humans, diseases are categorized as those that cause pain, distress, social difficulties, or even death. It is worth understanding the notion that diseases affect a person not only physically but also emotionally. When a person dies as a result of a disease, it is considered death by natural causes.

The primary categories of diseases are infectious diseases, non-infectious diseases, hereditary diseases, and physiological diseases. Infectious or communicable and non-infectious or noncommunicable diseases have two significant categories of health conditions, as illustrated by Kailash Nagar *et al*. (2022) Figure 1.1.

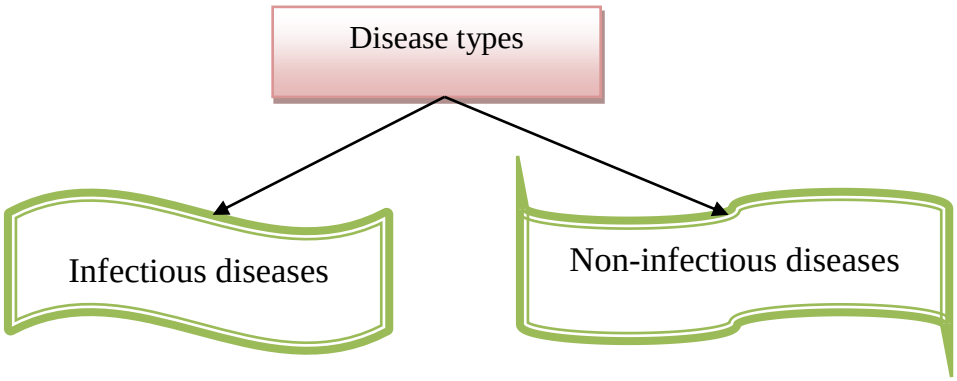

Figure 1.1 Types of disease

Infectious Diseases

Infectious diseases are spread from one person to another due to microorganic pathogens. These pathogens can exit from a host or infect a new person if the infected person emits body fluids. Usually, this kind of disease is considered airborne as it spreads through the air.

Non-infectious Diseases

Pathogens and other factors such as age, nutritional deficiency, etc, cause this type. Examples of this case can be hypertension, diabetes, and cancer; these diseases do not spread to others.

Infectious diseases have numerous symptoms and in severe cases it can damage human organs. Various vaccinations are distributed both inside and outside countries to curb the disease or at least mitigate its spreading property. World Health Organization also plays a significant role in spreading awareness and provide cautionary preventive measure for this disease.

It should be noted that all the above statements are made possible only with the help of medical professionals. Healthcare systems should be able to generate early and accurate disease forecast models in these situations, by Ireneous N Soyiri *et al*. (2012). Machine Learning (ML) based prediction techniques are developed to be applied for disease prediction. Disease prediction is the process of forecasting the progression of a medical condition in an individual by systematically analyzing the relevant health data. This raises the need for accurate and robust diagnostic systems for proper treatment. During the disease prediction phase, patients' health data is collected, and with the assistance of developed disease identification techniques, the presence of the disease is diagnosed. Figure 1.2 describes the process of recurring disease.

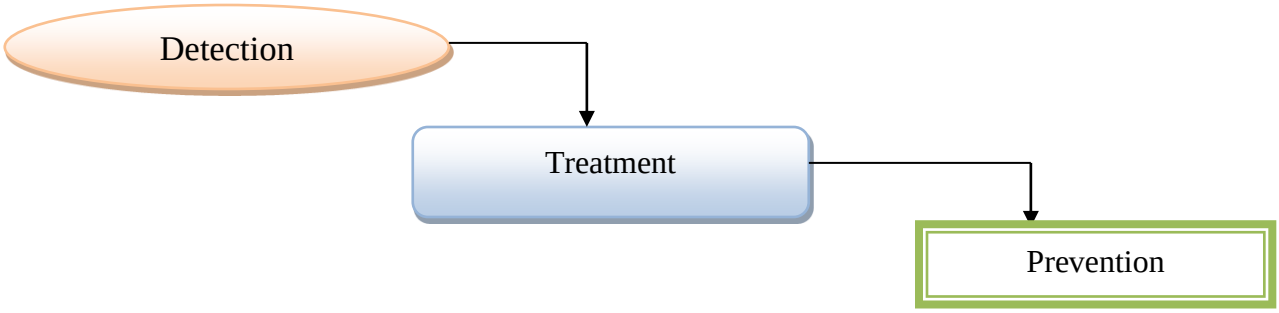

Figure 1.2 Recurring disease process

- **Prediction**

Disease prediction involves anticipating and monitoring the likelihood of disease conditions based on various influencing factors. Modern diagnostic techniques consider critical parameters such as age, sex, geographical location, and lifestyle habits. These approaches aim to provide accurate predictions through quantitative analysis while maintaining ease of use, ultimately supporting early detection and preventive healthcare strategies.

- **Treatment**

Treatment is when healthcare providers try to control the disease and its symptoms and slowly help patients recover. This procedure or therapy can involve medicine, therapy, surgery, or other strategies. The treatments that must be administered for infectious diseases are of great challenge, and the disease reservoirs help healthcare providers with crucial information that helps them move further.

- **Prevention**

Disease prevention involves stopping the spread of disease or controlling its effects. These are cost-effective and may be administered with or without a patient's disease.

1.2 COVID-19 AND PNEUMONIA DISEASE

1.2.1 COVID DISEASE

COVID-19 was agreed upon by Muhammad Ehsan Maqbool *et al*. (2024) to begin in Wuhan, China. Spreading around the globe and causing major catastrophes. The World Health Organisation reorganised COVID-19 as a communicable disease, and it was later determined to be contagious. During the initial stages, it was identified as SARS-CoV-2.

Symptoms of COVID -19 include,

- Fever
- Dry cough
- Dyspnea
- Headache
- Sore throat

- Rhinorrhea
- Fatigue
- Muscle pain
- In severe cases ARDS

COVID-19 is a transmissible illness that involves the respiratory system. Disease that targets this organ has a long-lasting effect. Hence, they must protect themselves and help others spread awareness of:

- Employ face mask
- Circumvent touching nose, mouth and ears
- Frequent hand washing with alcohol solutions as well as soap
- Avoid contact with contaminated people
- Maintain a reasonable distance from other people

These are the ways health care experts monitor and help infected people. During the quarantine period, in case of severity, in hospitals, patients are cared for according to the nature and severity of their symptoms and helpful self-management practices. Some of how they are taken care of are helpful treatment, intake of sufficient calories, and adequate water consumption to decrease the danger of dehydration.

1.2.2 PNEUMONIA DISEASE

Pneumonia, on the other hand, affects single or both lung sacs by fluid formation. The fluid causes cough due to phlegm, which gradually increases in severity. This disease is severe among infants, young children, old people over 65, and people with health issues.

General symptoms pneumonia is as given below,

- cough with phlegm green, yellow, or bloody in nature
- quick breathing as well as shortness of breath
- fast heartbeat
- fever, sweating, and chills

- fatigue
- nausea and vomiting

Pneumococcal conjugate vaccine and Pneumococcal polysaccharide vaccine are employed to prevent pneumonia. Treatment of the disease is based on its types—bacterial, viral, and fungal—and severity.

1.3 EARLY DISEASE PREDICTION

In the course of disease prediction, optimized machine learning and DL methods are employed by Samuel Darkwah *et al*. (2024) to identify the presence of disease and as well as to classify samples.

1.3.1 Data pre-processing

The disease dataset is collected from various hospitals and has errors. Through data mining, pre-processing models are used to eliminate abnormalities and convert raw data into numerical features effectively. Data normalization is another way to organize information within a database. Furthermore, this process helps protect the data, generates a more flexible database through redundancy, and removes conflicting dependencies. Therefore, data pre-processing comprises data selection, data normalisation, etc.

- Data collection - Is the process of data gathered from the given disease datasets.
- Data cleaning – Finding erroneous data and removes anomalies
- Data integration – A major factor of data management and fuses multiple sources into a single dataset.
- Data Transformation – Alters the data structure into numerical feature.
- Data normalizing – Converts the set of data values into a general scale.

1.3.2 Feature selection

The significant features of the pre-processed disease data are selected using an effective feature selection technique. Within the input data features, dimensionality reduction is applied to

attain better prediction performance. Feature selection embedded methods are used to remove irrelevant input data. Thus, it results in minimizing the computational load ratio as well as improving accuracy value with less consumption cost.

1.3.3 Classification

A practical, optimized classifier method is developed to classify the data features accurately. This classifier network accurately classifies whether the person is infected with the disease or not. The overall architecture of disease prediction is described in Figure.1.3.

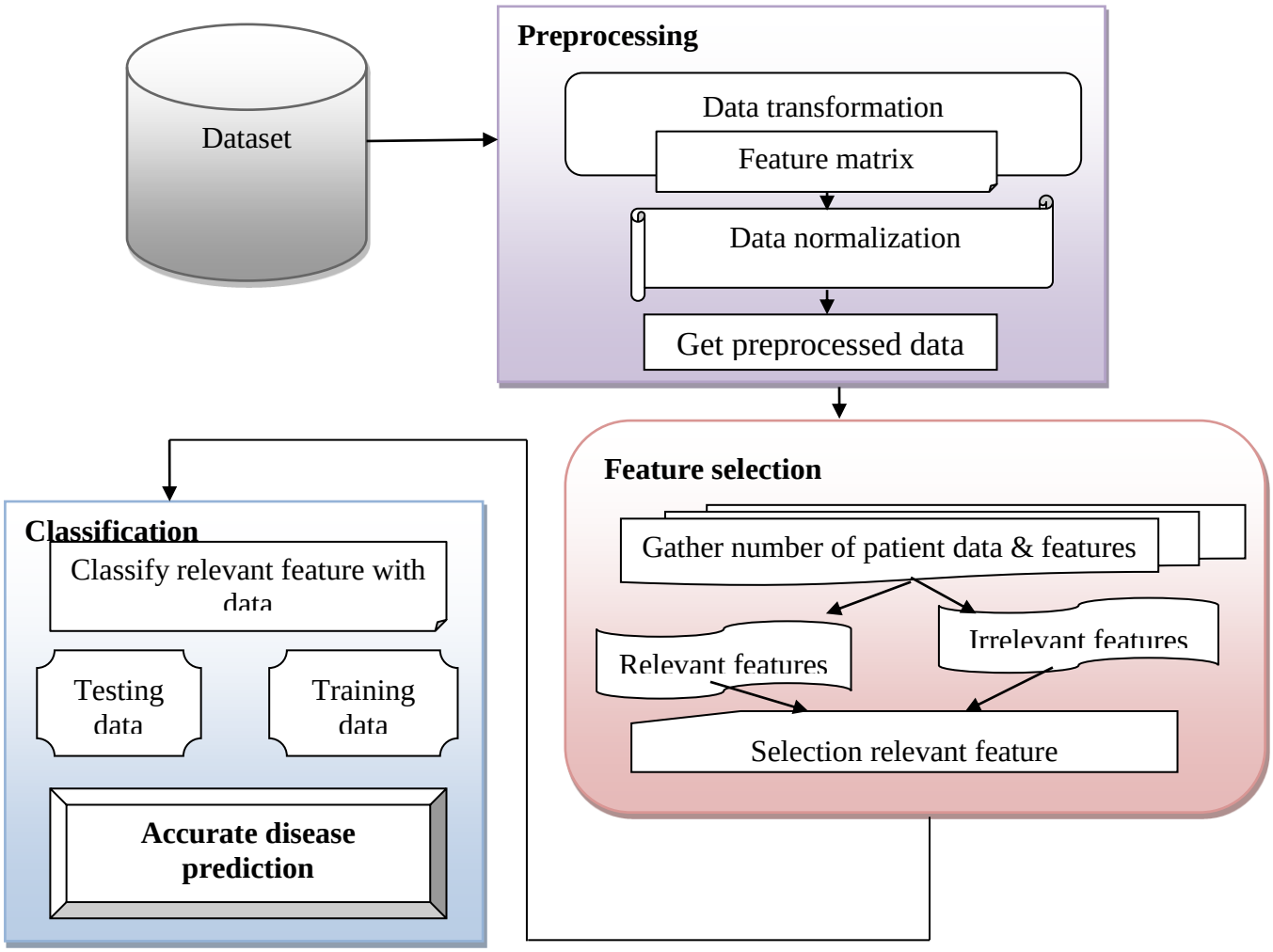

Figure 1.3 Overall Architecture Diagram of Disease Prediction

Figure 1.3 depicts the architecture diagram of disease prediction. Initially, patient information is collected from a given dataset. The sample input matrix is represented in terms of

patient data and features. For every sample, the data normalization is performed. Then, normalised data are transformed into a binary representation using an encoding technique. It obtains pre-processed data. It then performs a feature selection process for using this pre-processed diseased data with less time. An effective correlation method is used to discover the relevant and irrelevant features. Also, it performs statistical computation for finding well-matched features for illness prediction. Only it selects for relevant feature data. After feature selection, the patient data with extracted features is accurately classified. The testing and training data are picked in the dataset using the validation technique. Lastly, to get accurate patient data classification outcomes with less error and time.

Key Challenges

During the disease stage, numerous challenges are considered by Metty Paula et al. (2023) which are explained as below,

- Security and confidentiality
- Information distribution
- Data Correctness
- Patient collaboration

- Security and confidentiality

In the healthcare sector, data security and patient privacy are major issues, according to Xing Guo *et al*. (2022) . The patient's medical information is safely shared below particular conditions and for specific specialists. Thus, secured data communication methods are developed during this testing time of the disease, which calls for a fast solution.

- Information distribution

In the process of data extraction, data variety and volume play an essential role. For instance, COVID-19 , often leading to severe Pneumonia , has spread rapidly in all countries. So, there is a need to acquire precautionary measures to limit the spread of infection. So, developing an effective strategy for large-scale data secure sharing is extremely important.

- Data Correctness

Dates are transmitted through the internet and social media, which are places for presenting fake medical data and reports. Therefore, advanced data prediction is used to correct data presented on the internet. Some people look after the spread of wrong information. Therefore, fake information is taken down from the net.

- Patient collaboration

The patient should understand the newest disease attributes. Predictive systems should be made, and wearable sensors should collect physiological information. People should be aware of the significance of data sharing. In order to maintain data privacy, only authorized people should be allowed to collect the data.

The procedure of data mining advantages and Shortcomings are

Data mining advantages	**Data mining disadvantages**
❖ Prediction	❖ High-cost utilization
❖ Irregularity detection	❖ Protection
❖ Competitive expand	❖ Erroneous data
	❖ Violates user confidentiality

Applications of data mining

- Healthcare,
- Finance,
- Education,
- Retail, And Sports,
- Fraud Detection,
- Customer Segmentation,
- Market Basket Analysis, And
- Intrusion Detection.

Applications of machine learning for disease prediction

- Identification
- Intelligent system
- Decision and prediction

- Graph featurization
- Antibodies
- Trends and patterns
- Contact tracing
- Drugs and vaccination
- Screening and treatment

Applications of deep learning for disease prediction

- Medical imaging
- Disease tracking
- Protein structure forecast
- Drugs discovery
- Virus harshness as well as contagion

Healthcare ML applications

In identifying complicated patterns inside huge and triumphant information, ML methods are valuable, according to Virendra Kumar Verma and Savita Verma (2021). This ability presented by ML methods is particularly compatible with clinical relevance, especially for people who rely on higher genomics and proteomics measurements. ML methods are employed in diagnosing and predicting different diseases, making superior decisions on patients' recovery plans at medical relevance by executing a sound healthcare scheme. Hospital management employs this method to predict the stay times of patients waiting for a location in the department of interest. Clinics consider crisis room admissions by prognostic method. Therefore, ML execution may prove advantageous for patients, minimising costs and enhancing precision.

1.4 MOTIVATION

Disease prediction, including for conditions such as COVID-19 and Pneumonia, has been described from different viewpoints, and an enumerated assessment acts as a fundamental component in the computer relevance scheme. To evaluate the connection among enormous databases that are sustained as techniques to perform disease prediction, researchers develop new

techniques to come up with a novel method using mathematical formulas, geometric techniques, and software devices to meet the demand of the novel Disease.

Provoked by the speedy establishment for disease identification, this research is generated to enforce the Machine Learning and profound learning technique performance difficulties are resolved through programming language. The developed method carries data pre-processing, feature selection, and classification. Early disease prediction techniques are also generally employed within real-time applications. The existing research is concentrated on disease detection.

1.5 PROBLEM DEFINITION

Numerous diseases, like epidemics, tend to increase mortality rates in the healthcare domain. This is why people should be given proper treatment and prevented from spreading.

- Data pre-processing is essential in the disease prediction approach, and medical datasets have unbalanced, missing values. However, several existing data pre-processing techniques to improve forecast results fail to consider reducing space and time complexity during pre-processing.
- Countless Machine learning-based feature selection methods have been recently developed for disease prediction. However, these methods could not reduce the time involved in dataset pre-processing, and feature selection accuracy was not boosted.
- There is much research on early disease detection. However, efficient disease prediction classification accuracy and time are major demanding problems.
- This research work develops new and efficient pre-processing, feature selection, and classification techniques to overcome these problems.

Ahmed S. Salama *et al.* (2022) introduced a new technique of novel IoT and cloud-based blockchain model for detecting patients with the virus. It was able to control and reduce the spread of infection. Furthermore, it effectively managed the diverse people's conditions in the blockchain system by implementing smart contract rules accurately. This led to a high error rate.

Amgad Muneer *et al*. (2023) proposed a hybridization of GCN and GRU models for the mRNA deterioration field that detects the balance and mRNA sequence degradation hazard. The

paper also highlights the proposed model with GCN_CNN's efficiency and effectiveness. However, the designed models did not reduce the validation loss by a certain number of epochs.

A Hybrid Chi^2-MI-based feature selection model was developed by Samrat Kumar Dey *et al*. (2022) to identify chronic and non-chronic kidney disease. The designed model was better executed with an Extra Trees classifier and generated high accuracy. A machine learning-based model was designed and used during the early identification of chronic kidney disease and patient monitoring. However, it failed to develop a real-time diagnosis of kidney failure patients with the assistance of a web application.

Santosh Kumar *et al*. (2022) developed a method for incorporating multimodal deep learning techniques for early diagnosis. They extracted discriminatory features by fusing a chest X-ray-based model and a cough diagnostic model. These extracted patient features were accurately predicted through the weighted sum-rule fusion method. However, the techniques for selecting significant features were not highly accurate.

The DL technique depends on Long-Short-Term Memory (LSTM) introduced by Sourabh Shastri *et al.* (2021) to forecast disease. It also computes the proposed model's efficiency and error rate and fails to offer preventive measures for illness prediction cases.

The Advanced Hybrid Ensemble Gain Ratio Feature Selection (AHEG-FS) method was designed by Syed Javeed Pasha and E. Syed Mohamed (2022) for improved disease risk prediction. For discovering significant features, an ensemble learning feature selection technique was used. Then, the gain ratio feature selection method was implemented to rank the highest to the lowest number of cases based on prediction accuracy. The new feature reduction technique, Area Under Curve, was used to calculate the accuracy. In order to achieve the most effective feature subset, backwards feature elimination was applied to eliminate any less contributing features. However, the proposed model was not utilised in a variety of another disease database.

Novel Spider Monkey Based Generalised Intelligent (SMbGI) framework was designed by V. Laxmi Narasamma *et al.* (2022) for finding out the malware activities. It was to gather the patient's vaccine tweets from Twitter. Then, pre-processing is carried out among the redundant tweets from collected tweets, and to remove the aspect terms, feature extraction was executed.

Inside the data, malicious activities were also effectively analysed and identified through SMbGI framework. As well, the malware function was established in the system to verify the proposed model's effectiveness. Finally, it performs sentiment classification on the tweets, yet the prediction accuracy was not improved.

Xu L MagarR *et al.* (2022) discussed DL techniques to predict the number of COVID–19 cases. Furthermore, it improves significant developments during prediction performance. The fusion of DL methods effectively deals with disease data. Moreover, effective methods were developed for measuring the spread of disease, and they failed to attain a more accurate prediction outcome.

The Variant of Concerned Deep Learning (VOC-DL) prediction framework was developed by Zhifang Liao *et al*. (2022) to predict confirmed cases of the virus every day. The time line dataset consist of VOC variant data, which was processed by applying the slope feature method within the designed method. The designed framework was not performed for other variant prediction.

1.6 RESEARCH OBJECTIVES

This research is an attempt to overcome the problems in disease prediction. This led to the main objective of improving the performance of the disease prediction model for COVID-19 and Pneumonia using optimised Machine Learning and Deep Learning techniques. The secondary objectives of this research are

- To enhance data pre-processing for high accuracy, techniques are proposed that effectively transform and standardise data to improve model performance.
- To improve the prediction accuracy and select the relevant features using the proposed feature selection approaches.
- To enhance the prediction model with high accuracy and reduce error rate using machine learning and deep learning techniques.

1.7 RESEARCH CONTRIBUTION

The contribution of the research work is three-fold.

Concerning the first contribution, the Additive Log Ratio Transformed One Hot Encoding (ALRTOHE) Technique is proposed to execute data pre-processing with high accuracy in less time. The input for data is obtained from the COVID-19 and Pneumonia disease datasets. With this input data, pre-processing is carried out in the ALRTOHE technique via two dissimilar phases: additive log-ratio transformation and the one-hot encoding technique. In the initial phase, additive log-ratio transformation normalises data into a specific range. Secondly, data decoding transfers the numerical data into binary coding. Here, the data decoding is done to alter statistical data within binary coding. The propsed technique uses one-hot encoding to change numerical categorical variables into binary vectors. From here, binary representation of database is achieved in less time during data pre-processing. The proposed ALRTOHE technique outcome contrasts with conventional methods regarding pre-processing accuracy, space complexity and pre-processing time.

The Zero Mean Feature Normalised Encoding (ZMFNE) technique is proposed for highly accurate pre-processing of the input dataset. Also, it performs prepping of the data normalisation process and the encoding process. The redundant and conflicting data are removed during processing to reduce the complexity. The input data samples, as well as features, are acquired from the dataset. Then, the input data is normalised via zero-mean feature scaling. Afterwards, the data transformation process is executed using one-hot encoding. The encoder modifies the categorical features into a numeric array. The encoder obtains an integer array-like or a string as input. Here, features are encoded, and a binary column for each category is produced. From here, the data normalisation and the transformation process are performed, which aids in attaining the pre-processed data output. The proposed ZMFNE technique's performance analysis enhances pre-processing accuracy with minimal time.

The second contribution is a Nonlinear Sammon Projective Pattern Selection (NSPPS) Model proposed for choosing the relevant patterns with less error for disease prediction. This model utilises the patient data files as input. In the NSPPS model, Sammon projection projects the high to low dimensionality space to keep the inter-point distance structure. The Nonlinear Sammon Projection is used to pick the features observed from contagious disease by a newly

found illness. With the assistance of Sammon's mapping, distance of inter-pattern are maintained, and the error rate involved in choosing the pertinent features is reduced.

In order to pick the relevant features with high accuracy for disease prediction, the Tversky Similarity-Indexed Distributive Feature Embedding (TSIDFE) Technique is proposed. The designed model aims to cut down high-dimensional data to a low-dimensional space of analogous features. The pre-processed disease dataset is considered input for the TSIDFE technique. The input comprises several data points and their features. The TSIDFE technique is used to choose similar features from the input dataset. Then, the Tversky index similarity is utilized to measure the coefficient of two features in the TSIDFE technique. From that, the similar features are only used for disease prediction; otherwise, they are removed. This minimises the time and space complexity involved in the feature selection process.

Statistical correlative targeted projection pursuit-based feature selection (SCTPP-FS) Technique is proposed to perform high-accuracy pertinent feature selection with a lower error rate. In the SCTPP-FS technique, the feature selection process takes diverse features as input from the prepped database. By calculating the cohesion between the features, the target features are projected using Kaiser–Meyer–Olkin correlative projection pursuit. In turn, relevant features are picked with high accuracy in less time with a low error rate.

The third contribution proposes a novel Emphasis Perceptron Boosting Classification (EPBC) technique for disease prediction. This proposed EPBC technique uses input as patient data with selected features (patterns). With this input, the boosting algorithm assists in attaining accurate classification. With the help of the proposed EPBC, enhancing disease prediction accuracy in minimal time becomes a possibility. This intent initially builds a weak classifier through the weighted sum. The designed classifier classifies the patterns with nil training error. The weight and feature vector objective function are derived to offer a predicted outcome at an early stage. From that, the actual classification outcomes are acquired accurately.

A Time-dependent Cox regressive Levenberg–Marquardt Convolutional Neural Learning (TCLMCNL) Technique is proposed for disease prediction to attain accurate patient data classification. TCLMCNL Technique containsf the input layer, an output layer, and hidden layers. Several significant relevant features with data are considered input in the input layer. The

input is transferred to the hidden layer for the classification process. The time-dependent Cox regression is applied to calculate Cramér's phi correlation function in that layer. According to the regression outcomes, precise classification is executed. For each and every step, the Huber loss finds the predicted and actual outcomes. Additionally, the Levenberg–Marquardt algorithm diminishes error. Lastly, the output layer returns the prediction outcomes for disease detection. Thus, the prediction accuracy is augmented using the TCLMCNL Technique.

A Memetic Optimised U-Net Deep Learning (MO-UNetDL) classifier technique is developed for predicting disease via a classification process. MO-UNetDL technique comprises several layers for analysing patient data samples. The input layer takes assigned features from the dataset as input. After that, the weights are allocated to a set of input data samples inside the convolution layer to integrate the bias function. In order to find the similarity, Wilcox's index coefficient is used in this stage. The index coefficient is offered to the soft step activation in the hidden layer. The disease is accurately diagnosed if the activation function outcome is '1'. Subsequently, the data samples dimension is minimised by a max-pooling operation. Then, upsampling is performed with augment, data dimension, and finally, the classification outcomes are acquired. In order to reduce data classification loss, memetic optimisation is implemented in MO-UNetDL to tune the hyperparameters. To begin with, initialise the number of individuals and compute fitness. The top individual is picked through truncation selection. Two-point crossover generates new offspring. Through bit flip mutation, the input bit string is randomly exchanged. Once genetic operators are carried out, the picked individual is calculated, and its fitness is confirmed. Until an optimal solution is attained, this process is repeated. From this, the memetic optimised U-Net deep learning classifier identifies an optimal hyperparameter to lessen the error, resulting in higher disease prediction accuracy.

1.8 THESIS ORGANIZATION

This thesis includes an exploration of the various early disease prediction, a clarification of other work, and an overview of results and contributions. An overall framework of the thesis is given below:

Chapter 2 displays the literature review regarding disease prediction. It presents the dissimilar kinds of disease prediction of machine learning and deep learning models. A

dissimilar feature extraction method was utilised for features squeezing. Various classification algorithms are employed for better disease forecasting. The drawbacks faced by the existing disease prediction method are also discussed.

Chapter 3 portrays the research methodology. The proposed thesis executes pre-processing, feature selection, and classification techniques to achieve high accuracy with minimum time in disease prediction. Experimental analysis and performance evaluation of the different proposed techniques are developed to improve disease prediction using dissimilar metrics.

Chapter 4 portrays the ALRTOHE and ZMFNE techniques, which are carried out for data pre-processing with high accuracy and less computational complexity. In the ALRTOHE Technique, additive log-ratio transformation and the one-hot encoding technique are used to perform data normalisation and transformation. The proposed ALRTOHE and ZMFNE techniques increase the pre-processing accuracy and minimise time.

Chapter 5 describes the NSPPS Model, the TSIDFE Technique, and the SCTPP-FS Technique, which are proposed to select relevant patterns. In the NSPPS Model, Nonlinear Sammon Projection is used to pick the pertinent patterns with Sammon's mapping. In the TSIDFE technique, the Tversky index similarity coefficient is used to select relevant features. In the SCTPP-FS technique, Kaiser–Meyer–Olkin correlation projection pursuit is used to compute correlation for accurate disease prediction with greater accuracy, less time and error rate.

Chapter 6 explains the EPBC technique, TCLMCNL Technique, and MO-UNetDL classifier techniques, which are proposed for patient data classification. The EPBC technique, via a boosting algorithm, assists in performing accurate classification. The proposed TCLMCNL Technique predicts the disease. The MO-UNetDL classifier technique employs Wilcox's index coefficient and memetic optimisation to reduce data loss.

Chapter 7 portrays the ablation study for developing optimised machine learning and deep learning models. Each proposed contribution is examined for conducting an ablation study. The proposed research work, such as pre-processing, feature selection, and classification, was introduced for disease prediction with maximum accuracy and minimum time.

Chapter 8 summarises the analysis and evaluation of novel pre-processing, feature extraction, and classification techniques with a disease dataset. It introduces the application of AI in medical data for accurate prediction. The RSNA Pneumonia Detection Challenge database and COVID-19 database were employed to estimate the simulation using the proposed methods.

Chapter 9 describes performance analysis of machine learning and deep learning methods for disease prediction. It investigates classification without pre-processing and feature selection. Then, five different classification methods with pre-processing without feature selection are evaluated. It also examines classification with pre-processing and with feature selection.

Chapter 10 summarises the proposed approach's conclusion and outlines the potential future directions and scope of the designed approach.

1.9 CHAPTER SUMMARY

The above chapter discusses the important context of disease prediction. It briefly outlines the encouragement of this research and the research objectives. In addition, the problem statement explores contributions, the various phases of the proposed research methodology, and the organisation. The next chapter will present a review of the literature on the existing disease prediction models.