CHAPTER 5

NONLINEAR TVERSKY SIMILARITY AND STATISTICAL CORRELATIVE TARGETED PROJECTION BASED FEATURE SELECTION FOR DISEASE IDENTIFICATION

5.1 INTRODUCTION

The healthcare field involves real-time and accurate results. The detection and prediction of diseases are crucial to prevent fatal consequences. Early prediction of disease is essential, as it enables the effective quarantining of affected individuals and provides them with medical attention to curb its spread. It is not easy for doctors to physically recognise diseases, owing primarily to time constraints. The development of electronic medical records during recent years has been highly for the success of ML techniques. Many Machine Learning (ML) techniques have been developed to diagnose diseases, but the accuracy of these diagnoses has not improved. To improvise the accuracy of disease prediction, feature selection methods have been introduced.

Feature selection is carried prior to the prediction process to enhance the overall result of disease diagnosis. Feature selection is the task of recognizing meaningful and suitable aspects of the provided database that significantly increase the efficiency of the constructed models. The preliminary objective of feature selection is to decrease the number of features by selecting the most informative ones. Reducing the number of features offers several advantages to the practitioner during the analysis process. Additionally, the selection of features is crucial and provides further information to enhance the working efficiency of the developed models. It enhances the predictive power of these methods and reduces their training time. Therefore, feature selection is necessary to achieve more accurate prediction outcomes. Recently, numerous Machine Learning (ML)- based feature selection techniques have been developed for disease prediction. Yet, the accuracy of feature selection was not improved significantly to produce results with a minimum error rate. To solve these issues, three novel feature selection methods are developed in this research work.

The Nonlinear Sammon Projective Pattern Selection (NSPPS) method, a significant contribution to disease prediction, is proposed in this research. The NSPPS model, which uses

patient data files as input, employs Sammon projection to project from a high-dimensional space to a lower-dimensional space, maintaining the inter-point distance structure. The Nonlinear Sammon Projection is then used to select patterns that examine the disease as contagious, based on recently determined disease patterns. With the help of Sammon's mapping, inter-pattern distances are maintained, and the error rate is minimized in choosing pertinent patterns. This results in a powerful improvement in the accuracy and time efficiency of feature selection by the NSPPS model.

The Tversky Similarity Indexed Distributive Feature Embedding (TSIDFE) method, a novel approach to disease prediction, is proposed in this research. The TSIDFE technique is designed to map elevated-dimensional data to a minimum-dimensional space of analogous aspect. It takes preprocessed patient data as input and selects a minimum number of relevant features from various aspects. This innovative method promises higher accuracy in disease prediction. In the proposed TSIDFE technique, the Tversky index similarity coefficient between two features is calculated. The similarity coefficient varies from 0 to 1. If the similarity coefficient gives the result as '1', the feature is relevant; likewise if the result is '0', the feature is irrelevant. From the results, pertinent aspects are chosen for disease prediction. Irrelevant aspects are eliminated. This reduces the time and space complexity involved in the feature selection process.

Statistical correlative targeted projection pursuit-based feature selection (SCTPP-FS) Technique is proposed for pertinent feature selection with higher accuracy. The proposed SCTPP-FS technique is modeled to improve the accuracy of feature selection while minimising the error rate. The SCTPP-FS technique takes diverse features as input from a preprocessed dataset to perform the selection process. Then, the target features are projected by computing the correlation between the features. In the SCTPP-FS technique, the correlation between features is examined using Kaiser–Meyer–Olkin correlative projection pursuit to select the main features for accurate disease prediction. Through the correlation measure, relevant and irrelevant aspects are identified more efficiently. Based on this, relevant features are selected for their superiority and to minimise the error rate.

5.2 PROPOSED NONLINEAR SAMMON PROJECTIVE PATTERN SELECTION TECHNIQUE

Pattern identification is the task of mapping novel aspects to lesser aspects, which maintains leading data given in the database. Various kinds of pattern selection exist to choose the pertinent patterns that are current in the dataset. A novel NSPPS method is proposed to choose error-reduced pertinent patterns.

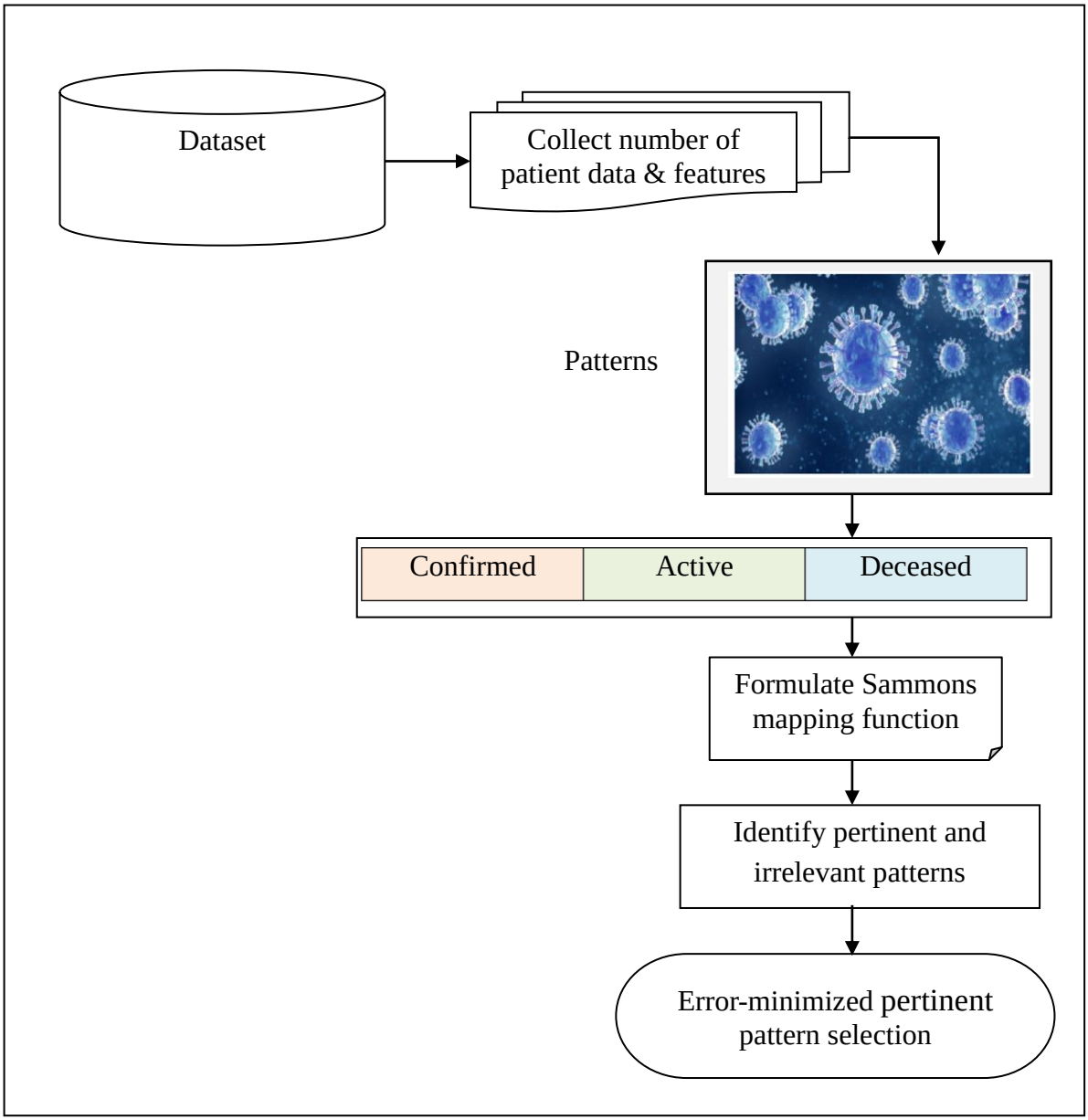

Figure 5.1 Structure of Nonlinear Sammon Projective Pattern Selection

Figure 5.1 demonstrates the structure of relevant pattern selection using Nonlinear Sammon Projection. As given in above Figure, a mapping function referred as 'fun' transforms a pattern 'Q' (patient files) with features $F = f_1, f_2, .., f_m$ where $m = 11$ of a 'n' input space to pattern 'P' of 'n' dimensional projected space (i.e., '$n < m$'). It is expressed as.

$$P=fun(Q) \qquad (5.1)$$

The equation 5.1 represents the projection of a high-dimensional input pattern Q in to a lower dimensional space. Q is the original pattern that is patient file with multiple features $f_1, f_2, .., f_m$ and fun is the Sammon mapping function, which is used for dimensionality reduction. Where P is the projected pattern in a new space of lower dimension n. The transformation aims to preserve important structural relationship between similarities in the reduced space.

In the above-mentioned equation (5.1), a criterion 'C' is introduced for optimal pertinent pattern selection. After which, the mapping 'f' is identified among all the transformations 'g' by meeting below condition.

$$C\{fun\ (Q)\} = \max C\ \{g(Q)\} \qquad (5.2)$$

The equation (5.2) defines how the optimal mapping function is selected. Where C is a selection criterion, that evaluates how well a projection captures relevant pattens. Among all possible mapping functions g, the function fun is selected as the one that maximizes the Criterion C.

The above provided mappings vary in functional forms of g and through a criterion which they optimize. Inter-pattern distances are maintained when using features 'F', and thereby optimized error-minimized pertinent pattern selection is achieved.

Next, the Nonlinear Sammon Projective Pattern Selection is carried out, using patient files as input to select the pertinent features for disease detection. Sammon projection is a nonlinear model that maps an elevated-dimensional space to a lower-dimensional space,

preserving the inter-point distance structure of the original space in the minimum-dimensional projection.

By managing distances between patterns under projection, conservation of this inherent inter-point distance structure is achieved for every patient file. Let us assume, pattern 'P_i' and 'P_j' is the inter-pattern distances at Input Space (IS) as well as their corresponding Projected Space (PS) is denoted as '$IS(P_i, P_j)$' and '$PS(P_i, P_j)$' respectively. The patterns in proposed NSPPS technique are considered as confirmed cases, active cases and deceased cases.

$$\frac{dC(t)}{dt} = \alpha\ (N) - \mu_1 A(t) \qquad (5.3)$$

From equation (5.3), confirmed cases at time 't' are acquired and depended on rate at that vulnerable population goes to long-established cases '$\alpha\ (N)$' as well as vulnerable population take onside as active people at time instance 't' referred as '$\mu_1(t)$'. Likewise, active cases at time 't' is computed as follows.

$$\frac{dA(t)}{dt} = \beta\ (N) - \mu_2(t) + \frac{dC(t)}{dt} \qquad (5.4)$$

Where '$dA(t)$' is computed on the basis of '$\beta\ (N)$' denoting the rate at which vulnerable population is transferred to active cases. As of long-established cases as well as rate at that confirmed cases '$\frac{dC(t)}{dt}$' are included to active '$\mu_2(t)$' correspondingly. In the same way, the deceased cases '$dD(t)$' are obtained as provided below.

$$\frac{dD(t)}{dt} = \gamma\ (N) \qquad (5.5)$$

The above equation (5.5), depends on rate at which active cases go to deceased cases '$\gamma\ (N)$', deceased cases are acquired. Depending on the above three patterns at consideration, patient file information is maintained after satisfying the condition given below through the Sammons mapping as follows.

$$E = \frac{1}{\sum_{P_i=1}^{m}\sum_{P_j=1}^{n} IS(P_i,P_j)} \sum_{P_i=1}^{m}\sum_{P_j=1}^{n} \frac{\left[IS(P_i,P_j) - PS(P_i,P_j)\right]^2}{IS(P_i,P_j)} \qquad (5.6)$$

From the above equation (5.6), with the aid of the Sammons mapping, error is said to be reduced while performing pattern selection via estimating the distance between the '$i-th$' pattern and '$j-th$' pattern, in the original input space '$IS(P_i, P_j)$' and the distance between their projected space '$PS(P_i, P_j)$' respectively.

The following Algorithm 5.1 depicts the Nonlinear Sammon Projective Pattern Selection process. The designed algorithm is developed with the aim of selecting optimal and error-minimised relevant features. To achieve this, three criteria are considered, and patterns are identified as infectious diseases. The recently revealed disease was chosen effectively via a nonlinear Sammon Projection. With the help of Sammon's mapping, inter-pattern distances are managed and thus decrease the error rate involved in choosing pertinent patterns.

//**Algorithm 5.1**: Nonlinear Sammon Projective Pattern Selection
Input: Dataset 'DS', district level counts 'DLC', Patient-wise data 'PD', patient number 'PID' **Output**: Optimal and error-minimized relevant pattern selection 'RP'
step 1: **Initialize** time instance 't' **step 2**: **Initialize** rate at which susceptible population goes to confirmed cases 'α', susceptible population take onside as active people 'μ_1' **step 3**: **Initialize** rate at which susceptible population goes to active cases from confirmed cases 'β', rate at which confirmed cases are added to active cases '$\mu_2(t)$' **step 4**: **Initialize** the rate at which active cases goes to the deceased cases '$\gamma\ (N)$' **step 5**: **Begin** **step 6**: **For** each Dataset 'DS' with district level counts 'DLC' obtained from Patient-wise data 'PD'patient files with patient number 'PID' **step 7**: Formulate a mapping function 'fun' as defined in Equation (5.1), such that it maximizes criterion C as specified in Equation (5.2) **step 8**: **For each** time instance 't' **step 9**: Estimate confirmed cases as in equation (5.3) **step 10**: Estimate active cases as in equation (5.4)

step 11: Estimate deceased cases as in equation (5.5)

step 12: **End for**

step 13: Formulate Sammons mapping as in equation (5.6)

step 14: **Return** (relevant patterns 'RP')

step 15: **End for**

step 16: **End**

5.3 PROPOSED TVERSKY SIMILARITY INDEXED DISTRIBUTIVE FEATURE EMBEDDING (TSIDFE) TECHNIQUE

The Tversky Similarity-Indexed Distributive Feature Embedding (TSIDFE) technique is proposed for selecting similar features with higher accuracy in prediction. Feature selection helps reduce the computational cost of modelling, thereby enhancing the model's efficiency. The dataset typically includes numerous features that require more time during the learning process, which can cause the classifier to lack sufficient information while performing an exact classification task. It also degrades the accuracy metric. Hence, the necessary value for selecting pertinent features is to reduce complexity and enhance classification accuracy.

Samrat Kumar Dey *et al.* (2022) developed a hybrid Chi^2-MI-based feature selection model to diagnose chronic diseases. Chi^2-MI removes the superfluous aspect, and the Pearson correlation matrix was employed to select the top significant aspects for the forecast. However, the accuracy of feature selection was not improved with minimum error. To overcome these problems, the proposed TSIDFE performs the relevant feature selection with higher accuracy.

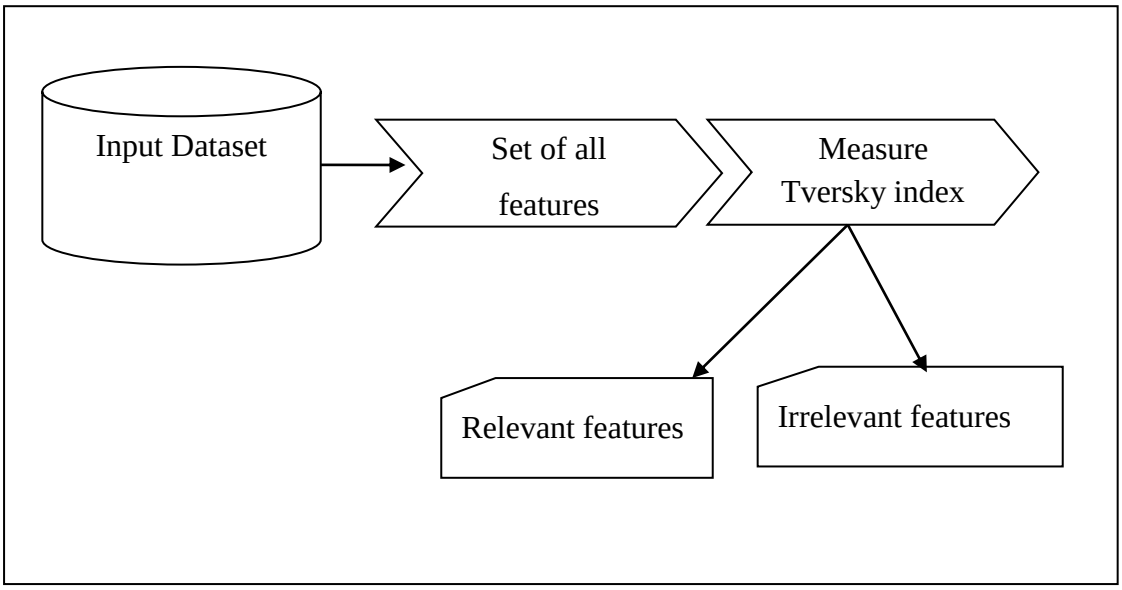

Figure 5.2 Tversky Similarity Indexed Distributive Feature Embedding Technique (TSIDE)

Figure 5.2 shows the process of the TSIDFE technique to accurately select the features in less time. The proposed technique is a low-dimension projection method that aids in projecting high-dimensional data into a lower-dimensional space of comparable features, described by high similarity, and dissimilar features are described with minimal similarity.

Let's assume a set of features allocated in high-dimensional space $F = f_1, f_2, .., f_m$. Tversky index is used to measure the similarity between features using equation 5.7.

$$\delta = \frac{[f_i \cap f_j]}{K(f_i \cap f_j) + L(f_i - f_j)} \qquad (5.7)$$

Where 'δ' point outs a similarity coefficient, 'f_i' and 'f_j' point outs two features, '$f_i \cap f_j$' refers to mutual dependence among two features, '$f_i - f_j$' refers variance among two features. From equation (5.7), 'K' and 'L' refers to metrics of Tversky index ($K, L \geq 0$). Similarity Coefficient (δ) gives output results among [0, 1]. Based on the coefficient outcomes, more similar features are properly detected, as mathematically demonstrated below.

$$\delta = \begin{cases} \beta = 1,\ relevantfeatures \\ \beta = 0,\ irrelevantfeatures \end{cases} \qquad (5.8)$$

From the above equation (5.8), pertinent aspects are selected for accurate disease detection and irrelevant aspects are eradicated. Based on chosen features, disease detection is carried out with minimum time.

//Algorithm 5.2: Tversky similarity indexed stochastic distributive feature embedding technique
Input: Dataset, preprocessed features $F = f_1 f_2, .., f_m$ **Output:** Select relevant features
Begin **step 1** : Collect the number of features $\{F = f_1 f_2, .., f_m\}$ **step 2 : For each** feature 'f_i' and 'f_j' **step 3 :** Measure the similarity '$\beta\ (k_i, k_j)$' **step 4: if** $(\delta = 1)$**then** **step 5:** The feature is said to be relevant **step 6: Select** relevant features **step 7: else** **step 8:** The feature is said to be irrelevant **step 9: Remove ir**relevant features **step 10: End if** **step 11: End for** **End**

Algorithm 5.2 explains the process of pertinent feature selection depending on the similarity measures using the TSIDFE technique. Initially, features are collected from the input dataset. After that, the similarity index is measured to discover the pertinent aspects. When the similarity coefficient gives '1', feature is considered as relevant. Or else the feature is irrelevant. Similar features are selected for disease forecasting, whereas further features are eliminated. As a result, accurate feature selection is achieved in less time and with less space complexity.

5.4 PROPOSED STATISTICAL CORRELATIVE TARGETED PROJECTION PURSUIT-BASED FEATURE SELECTION

Feature identification is performed to make disease forecasting as less complicated. Through a greater number of features, learning accuracy and training speed are significantly affected. To address these issues, only essential features are considered for enhancing the accuracy of disease forecasts in a shorter timeframe. Syed Javeed Pasha *et al.* (2022) designed the AHEG-FS method for improving the detection of this disease, which is categorised as dangerous; however, the space complexity was not minimised. Therefore, Statistical Correlative Targeted Projection Pursuit-based Feature Selection (SCTPP-FS) is developed to accurately learn from patient data in the dataset by selecting more informative features with minimal memory consumption.

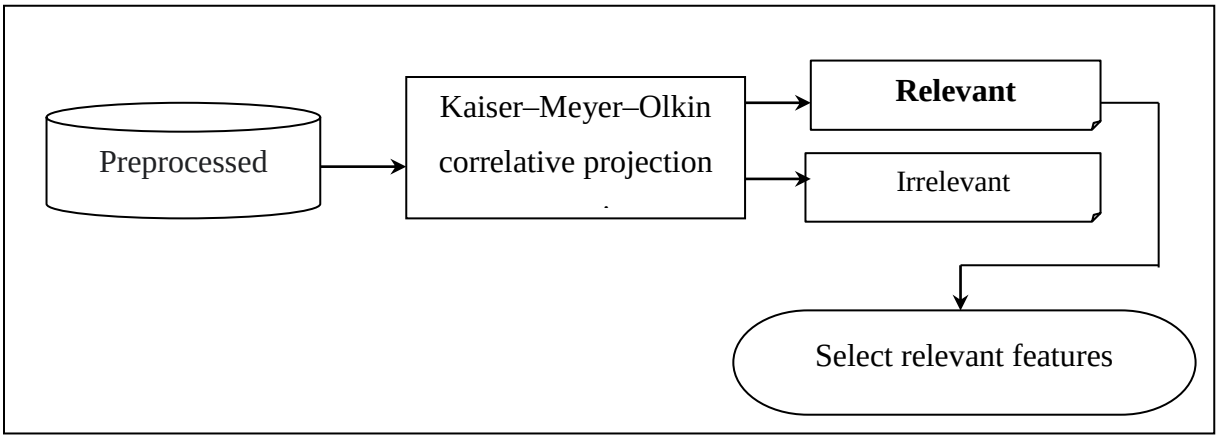

Figure 5.3 Process of Statistical Correlative Targeted Projection Pursuit-based Feature Selection

Figure 5.3 illustrates the process of the SCTPP-FS technique for selecting additional pertinent features from the database. The designed SCTPP-FS takes a preprocessed dataset as input. With the input, a Kaiser–Meyer–Olkin (KMO) correlation-based projection pursuit is employed to identify the relevant and irrelevant features. The Targeted projection pursuit is a type of dimensionality reduction model used to select features. It permits us to interactively discover extremely intricate data and identify the significant features using the Kaiser–Meyer–Olkin correlation. KMO is a mathematical measure used to determine well-matched features for

disease prediction. The feature analysis is conducted to identify the correlation between the two features, which is presented below.

$$\rho = \sum_{i=1}^{n} \sum_{j=1}^{m} C\left(f_i, f_j\right) \quad (5.9)$$

Where 'ρ' symbolizes a correlation coefficient, which represents overall relationship between feature pairs. f_i, f_j represents features from the dataset. i=1 to n means looping over n features in one group. j=1 to m means looping over m features in another group. $C(f_i, f_j)$ is the correlation measure between feature f_i and feature f_j.

$$C\left(f_i, f_j\right) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} |f_i - f_j|^2}{\sum_{i=1}^{n} \sum_{j=1}^{m} |f_i - f_j|^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} {S_{ij}}^2} \quad (5.10)$$

Where $\left|f_i - f_j\right|^2$, computes total squared difference between each pair of features $f_i \; and \; f_j$, which is used to capture how different two features are from each other.

$$S_{ij} = \frac{|f_i - f_j|}{\sqrt{(1-f_i)^2 \left(1-f_j\right)^2}} \quad (5.11)$$

'S_{ij}' indicates partial correlation amid features. Also it defines how close the feature values are to 1. The results of correlation coefficient gives output between the 0 and 1.

$$\rho = \begin{cases} 1, & relevant \; features \\ 0, & irrelevant \quad features \end{cases} \quad (5.12)$$

In equation (5.12), the correlation coefficient 'ρ' provides '1' denotes the relevant features whereas, '0' denotes features are not pertinent. Pertinent features are identified as target features and it is projected for precise illness forecast as of elevated dimensional space to minimum dimensional space (L) in a minimum period.

$$M: T_f \; \rightarrow \; D_L \quad (5.13)$$

Where, 'M' refers to a projection function used for dimensionality reduction, 'T_f' refers the space of target feature (i.e. pertinent feature) mapped into a low dimensional space 'D_L' after

projection. So, the function M maps each feature vector in the high dimensional space T_f to a lower dimensional representation in D_L.

To enable faster and more accurate disease prediction, the relevant features are selected and projected from a high dimensional projected space to lower dimensional space using mapping function M. This dimensionality reduction helps eliminate redundancy, reduce computational complexity, and enhance the precision of the proposed model.

// Algorithm 5.3: Statistical correlative targeted projection pursuit-based feature selection

Input: Preprocessed big dataset, number of features $F \in f_1, f_2, \dots, f_n$

Output: Select significant features for disease prediction

Begin

step 1: Collect the number of features $F \in f_1, f_2, \dots, f_n$ from pre-processed dataset

step 2: **For** each feature f_i and f_j

step 3: Determine the correlation between features '$\rho = \sum_{i=1}^{n} \sum_{j=1}^{m} C\left(f_i, f_j\right)$'

step 4: **if** $(\rho = 1)$ **then**

step 5: Features are identified as relevant

step 6: **else**

step 7: Features are identified as irrelevant

step 8: **end if**

step 9: Project the target features into low-dimensional space '$M: T_f \rightarrow L$'

step 10: Remove the other irrelevant features

step 11: **Return (target features)**

step 12: **End for**

End

Algorithm 5.3 describes the steps involved in the feature selection using statistical correlative targeted projection pursuit. Initially, several aspects are considered as preprocessed databases. Following this, target features are mapped by computing the association among features. The feature with the highest correlation is chosen as the target aspect to predict disease accuracy.

5.5 EXPERIMENTAL SETUP

Experimental analysis of three proposed NSPPS Models, TSIDFE Technique and SCTPP-FS Technique, are performed with the implementation of Python. To investigate the result of feature selection, the COVID-19 coronavirus India database and RSNA Pneumonia datasets are used. In this work, the 10 cross-validation is employed. Overall, the database is categorised into a training and testing set, where 80% of the patient data is used for training and 20% of the information is used for testing. 80:20 splitting ratio is best for the provided database. From the dataset, 10000 to 100000 data samples are used as input for conducting the experiments. The results of the three proposed techniques are compared with those of the conventional AHEG-FS model and the hybrid Chi^2 and MI feature selection model, referred to as Chi^2-MI. The testing metrics used for analysing the performance of proposed and existing methods are presented as follows.

- Feature selection accuracy
- Feature selection time
- Space complexity
- Error rate

Table 5.1 Overall Features in COVID-19 and Pneumonia Datasets

Features Name (COVID-19)	Features Name (Pneumonia)
Date	Patient Id
Name of State / UT	Boxes
Latitude	Target
Longitude	Class
Total Confirmed Cases	Lung Opacity
Death	No Lung Opacity/ Not Normal
Cured / Discharged / Migrated	Normal
New cases	Lung Opacity
New deaths	No Lung Opacity / Not Normal
New recovered	Normal

Table 5.2 Selected Features in COVID-19 dataset using Proposed Feature Selection Techniques

Existing AHEG-FS	**Feature Name (NSPPS Technique)**	**TSIDFE Technique**	**SCTPP – FS Technique**
Date	Date	Latitude	Latitude
Name of State / UT	Latitude	Longitude	Longitude
Latitude	Longitude	Total Confirmed Cases	Cured/Discharged/Migrated
Longitude	Total Confirmed Cases	Death	New Cases
Total Confirmed Cases	Death	Cured/Discharged/Migrated	New Deaths
Death	Cured/Discharged/Migrated	New Cases	
Cured / Discharged / Migrated	New Cases	New Deaths	
New cases	New Deaths		
New deaths			

Table 5.3 Selected Features in Pneumonia dataset using Proposed Feature Selection Techniques

Existing AHEG-FS	**Feature Name (NSPPS Technique)**	**TSIDFE Technique**	**SCTPP – FS Technique**
Patient Id	Patient Id	Patient Id	Patient Id
Boxes	Boxes	Boxes	Boxes
Target	Target	Target	Target
Class	Class	Class	Class
Lung Opacity	Lung Opacity	Lung Opacity	Lung Opacity/ Not Normal
No Lung Opacity/ Not Normal	No Lung Opacity/ Not Normal	No Lung Opacity/Not Normal	Lung opacity
Normal	Lung Opacity/ Not Normal	Lung Opacity/ Not Normal	
Lung Opacity	Normal		
No Lung Opacity / Not Normal			
Normal			

Table 5.1 shows the overall features in the COVID-19 and Pneumonia datasets and compared with existing features. Table 5.2 lists the selected features in the COVID-19 dataset using the existing and proposed techniques, where in the NSPPS model, the features are reduced to 8, and while using TSIDFE model again, the features are reduced to 7, followed by SCTPP-FS model where the features are again reduced as 5. Table 5.3 lists the features of the Pneumonia dataset using the proposed techniques. In the NSPPS model, the selected features are 8, while using the TSIDFE model, the features are reduced to 7, and finally, in the SCTPP-FS model, the features are reduced to 6.

5.6 CHAPTER SUMMARY

A novel and efficient feature selection technique, the NSPPS Model, TSIDFE Technique, and SCTPP-FS Technique, is proposed for selecting relevant features to predict disease. The comparison analysis of the proposed NSPPS Model, TSIDFE Technique, and SCTPP-FS Technique is made with the existing AHEG-FS and Chi^2-MI methods. As for experimental results, it is clear that the proposed NSPPS method, the TSIDFE Technique, and the SCTPP-FS Technique efficiently perform feature selection for disease detection. The performance of the proposed NSPPS Model, TSIDFE Technique, and SCTPP-FS Technique are verified with the help of feature selection accuracy, feature selection time, error rate and space complexity. This helps in Attaining outcomes that verify the SCTPP-FS Technique is considered the best for accurately selecting the more pertinent features compared to other feature selection methods, From the Table 9.2 (a) and (b) NSPPS Model achieves 86% and 89%, TSIDFE Technique achieves 89% and 91%, AHEG-FS accuracy as 84% and 87%, and Chi^2-MI method as 81% and 85% for COVID-19, Pneumonia datasets in that order. The investigation results of the proposed SCTPP-FS Technique feature selection accuracy as 91% and 94%, with lower space complexity, less in time and error rate, in compare with existing techniques. The next chapter presents the intended classification techniques with experiments results.