Journal of Theoretical and Applied Information Technology
31st May 2022. Vol.100. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EARLY COVID-19 SPREAD PREDICTION USING SAMMON PROJECTIVE AND PERCEPTRON BOOSTING

KALAISELVI[1*], VIJAYABHANU[2]

[1] Research Scholar,Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education, Coimbatore, India.
[1]Asst. Professor, Department of Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore, India.
[2]Asst. Professor(SG), Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education, Coimbatore, India.

E-mail: [1]srkalai2010@gmail.com, [2]vijayabhanu_cs@avinuty.ac.in

ABSTRACT

Text mining is an Artificial Intelligence (AI) technology with Natural Language Processing (NLP) that converts unstructured text and databases into meaningful information. The prediction of disease at earlier stage is an essential and demanding task. Numerous existing methods were introduced for performing efficient disease prediction at an early stage. However, with the inception of new strains the accuracy and time with which the prediction was made was not found to be satisfactory. In order to address these issues, Nonlinear Sammon Projective and Perceptron Boosting (NSP-PB) method for early COVID-19 prediction is introduced. First, a Nonlinear Sammon Projective Pattern Selection is applied to the input patient files to select relevant patterns. Next, Emphasis Perceptron Boosting Classification is applied to the selected relevant pattern to categorize the COVID disease patient files. Here, Emphasis Boost Classifier merges weak learner result to form strong classifier output. This in turn helps to improve COVID disease prediction with higher accuracy and minimal time consumption. Experimental evaluation is carried out for factors such as prediction accuracy, prediction time and error rate with respect to number of patient files. The experimental result reveal that the proposed NSP-PB performs better with a 9% improvement in prediction accuracy, 39% reduction of error rate, and28% faster prediction time for Covid 19 spread prediction compared to existing works.

Keywords: *Artificial Intelligence, Natural Language Processing, World Health Organization, Nonlinear Sammon, Emphasis Perceptron, Boosting Classification.*

1. INTRODUCTION

A new machine learning forecasting model was introduced] to forecast the COVID-2019 spread [1]. Here, linear regression, multilayer perceptron and vector autoregression models were employed for COVID-19 data with the purpose of predicting the epidemiological instance of ailment and pace of COVID-2019 cases in India.

Despite efficient prediction, the prediction time consumed was not focused. "A new hybrid feature representation method was introduced by integrating hybridized random forest and hybrid features with the purpose of forecasting the antioxidant protein level [2]. Here, feature extraction and hybrid feature representation were employed to perform classification. Though classification was ensured, however the feature selection was not performed in an accurate fashion".

Cardiac disease prediction assisted the practitioners for making accurate decisions with respect to the concerned health of the patient. However, with the involvement of vast amount of features, training time was said to be compromised. "A dimensionality reduction method was introduced for identifying the relevant heart disease features via feature selection model [3]. Though space complexity was minimized, prediction accuracy was not ensured".

"Yet another machine learning model was introduced with the objective of predicting the number of upcoming patients caused by COVID-19 [4]. Here, four forecasting techniques, namely linear regression (LR), least absolute

shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) were applied to predict the threatening COVID-19 factors". ES performed better in forecasting the new confirmed cases, death rate and recovery rate. But, the computational cost was not minimized by existing forecasting techniques.

The research study was focused on the early stage prediction of disease and it is challenging task in COVID 19 disease. Many traditional approaches were developed to perform efficient disease prediction at an early stage. But, the accuracy was not sufficient and time was not reduced. Also, the feature selection was not performed and the error rate was not reduced. In order to overcome the issue, the paper focuses to pervade the void of the conventional healthcare system, using machine learning (ML) algorithms to simultaneously process healthcare and analyze data along with other features of COVID-19 patients, to predict the most likely outcome of a patient, based on date on which case is announced, city/district in which case is detected, contracted from which patient, nationality, type of transmission, source 1, source 2, source 3 etc. Current data science and artificial intelligence approaches are used for COVID-19 pandemic. IT/Computer scientists are applied to better detect, diagnose, treat and prevent the spread of the deadly virus. IT/Computer science applications are important to focus the prediction of disease issue in early stage. The main contribution of the proposed framework is presented below,

- To find COVID infected person in early stage, the machine learning boosting algorithms are used to provide processing of healthcare and patient data in place of the conventional healthcare system.
- To select error-minimized optimal and relevant patterns as compared state-of-the-art methods, Nonlinear Sammon Projective Pattern Selection is applied that are available for processing patient data.
- To achieve the accurate classification result, Emphasis Perceptron Boosting Classification algorithm is employed to fine-tune the parameters of the Boosted Perceptron.
- The proposed NSP-PB is implemented in Python then the performance of the proposed framework is compared with the existing techniques. The experimental results show that not only has shows minimum error rate but also achieves better performance in terms of prediction accuracy, prediction time than conventional COVID 19 prediction methods.

The rest of the article is organized as follows: In Section 2, related work reviews the articles discussed by different authors using state-of-the-art methods followed by the detailed discussion of the proposed materials and methodology used in detail in Section 3, along with the dataset description, pattern selection model and data analysis of the prediction algorithm used. Section 4 provides the data set description and the experimental setup being utilized. Section 5 discusses the result of the experiment by making comparison with the state-of-the-art methods. Finally, concluding remarks are provided in Section 6.

2. RELATED WORKS

The coronavirus disease (COVID-19) is represented as a denoted to as a severe acute respiratory syndrome coronavirus 2 SARS-CoV-2. It was first identified in Wuhan, China, in December 2019. Upon comparison with Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), COVID-19 is said to be transmitted swiftly by means of contact human and infected surfaces. Moreover, COVID-19 is said to be highly dangerous for patients having the condition of weak immune systems, the elderly or the senior citizen and patients with chronic disease. Also the governments in a global manner modeled several mechanisms to proceed with intensive measures in several key facets to monitor the crisis of COVID-19 outbreaks.

"Chie-square and Principal Component Analysis was applied for predicting heart disease [3]. By applying dimensionality reduction and identifying key features using feature selection, accurate decisions regarding patient health were said to be made. Four standard forecasting methods, called, linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) were analyzed to predict the threatening aspects of COVID-19 [4]. Disruptive technologies for analyzing COVID 19 using multiple criteria decision making methods was applied to perform effective diagnosis" [5].

The current outbreak of corona viruses is reported in Wuhan, China during late December 2019, followed by which on January 30, 2020, the world health organization (WHO) declared it as the Public Health Emergency of International Concern (PHEIC) with the spread of these wave more than 18 countries. In recent years, several research articles have been published on corona virus.

"A Novel Feature Reduction (NFR) model integrating Machine Learning (ML) and Data Mining (DM) algorithms were proposed to minimize the error rate and to accelerate the performance rate [6]. The NRF method used two different methods to achieve a robust and efficient disease risk prediction. The first method was designed on the basis of the heuristic process by minimizing the features requirement for further processing and then the second method acquired all the individual features for ensuring accuracy, therefore reducing running time to a greater extent. Machine learning approach was employed with the purpose of extracting the activities and trends of corona virus related issues" [7].

"A holistic approach of the knowledge of the epidemiological features of this corona virus is critical to control its spreading. A step wise model via cleaning, feature extraction and classification was proposed [8].Artificial intelligence-based were applied with the purpose of predicting the pervasiveness of the outbreak of COVID-19 in Egypt" [9].

The significance of this novel corona virus is said to be tremendously communicable respiratory disease is that it occurs in both the form of symptomatic and asymptomatic patterns in those patients who are already being infected, thereby resulting in a rapid increase in the number of contractions of both the confirmed and active cases. It is, hence found to be critical to accelerate the procedure of early detection and diagnosis of disease globally.

"Yet another feature integration method combining matrix-based representation and convolutional neural networks (CNN) was proposed for extracting the feature and conducting fusion to exceptionally utilizing the multi-source data structure to aid in medical decision making in an effective manner [10]. A case-based reasoning framework for early detection and diagnosis of novel corona virus utilizing semantic-based mathematical modeling was proposed therefore reducing the false positive rate of diagnosed case [11]. A consideration and review of immune-related associations in corona virus was investigated [12]".

"A smart health monitoring for heart disease prediction using ensemble deep learning and feature fusion to improve the accuracy involved in predicting was proposed [13]. A method for COVID-19 time series prediction globally employing a hybrid ensemble modular neural network that integrates nonlinear autoregressive neural networks was proposed [14]". The method eliminated irrelevant and redundant features, therefore minimizing the burden involved in prediction to a greater extent.

"A machine learning approach of COVID-19 diagnosis based on symptoms was proposed [15]. A review of artificial intelligence for COVID-19 containment was investigated [16]. A review of machine learning and deep learning methods for analyzing and detecting COVID-19 was investigated [17]".

"Outbreak prediction of COVID-19 for two extremely different dense and population countries employing machine learning was proposed [18]. The role of machine learning algorithms employing machine learning techniques for improving testing accuracy involved in predicted was designed [19]".

A novel IoT and Cloud based Blockchain Model was introduced for identifying the covid 19 patients [21]. But, the error rate was higher. Deep Learning Based LSTM Models were designed to accurately find the abnormal activity prediction [22]. However, the prediction time was not reduced. Optimized Deep Convolutional Neural Network with Cuckoo Search (DCNN-CS) Algorithm was designed for categorizing leaf diseases [23]. An in-ship localization algorithm was introduced for obtaining the higher accuracy [24]. A new Spider Monkey based Generalized Intelligent (SMbGI) framework was developed for discovering the malware activities [25]. But, the prediction accuracy was not improved. 3D Convolutional Neural Networks were introduced to achieve earlier and exact outcomes [26]. The designed network prediction time was higher.

In spite of the uncertainty connected with medical predictions, nevertheless predicting is essential in permitting us to better comprehend the prevailing hypothesis and analyze for the future accordingly. In this paper, Nonlinear Sammon Projective and

Perceptron Boosting (NSP-PB) method for early COVID-19 prediction is proposed to provide statistical forecasts and therefore early prediction for the Confirmed, Active, Deceased and Recovered model of COVID-19 using machine learning is presented.

3. PROPOSED NONLINEAR SAMMON PROJECTIVE AND PERCEPTRON BOOSTING (NSP-PB) METHOD

The novel coronavirus disease 2019 (COVID-19) pandemic generated by the SARS-CoV-2 pick up the threads to give rise to a critical and emergency menace to health globally. The epidemic in early December 2019 in the Hubei province has proliferated globally. This pandemic pursues to confront medical line systems widespread in several facets, together with intense shoot up in insistences for hospital beds and critical shortages in medical equipment, while several healthcare workers have personally been affected. In this paper we propose a Nonlinear Sammon Projective and Perceptron Boosting (NSP-PB) method for early COVID-19 prediction. Figure 1 shows the structure of NSP-PB method for early COVID-19 prediction.

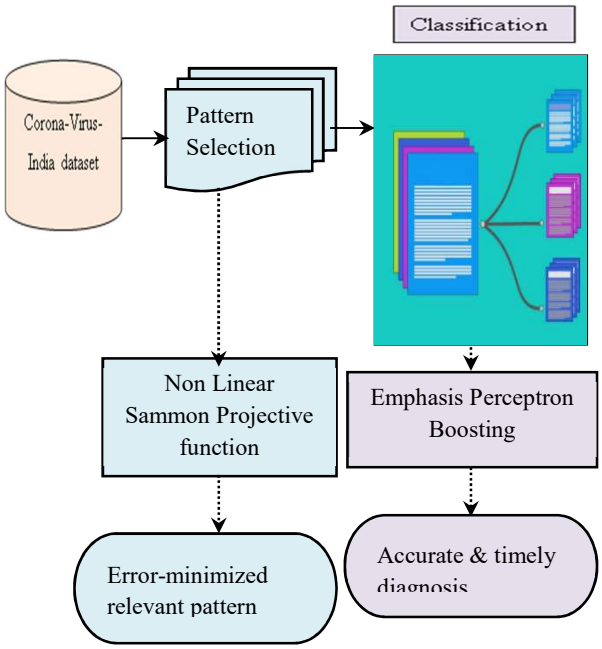

Figure 1: Structure of NSP-PB method

As shown in the above figure, two distinct processes are carried out, namely, pattern selection and classification. First, relevant pattern selection is made by applying the Nonlinear Sammon Projective function to the input patient files acquired from the corona-virus-india dataset [20]. With the acquired relevant pattern, classification is made by employing the Emphasis Perceptron Boosting with which accurate and timely diagnosis are done, therefore contributing to minimum mortality rate. The elaborate description of the proposed NSP-PB method is provided in brief in the following sections.

3.1 Nonlinear Sammon Projective Pattern Selection – minimizes error rate

Pattern selection is the process of mapping original features into fewer features that preserve the foremost information present in the dataset. Different types of pattern selection are said to prevail for selecting the relevant pattern present in the corona virus patient. In our work, a Nonlinear Sammon Projective Pattern Selection model is employed for selecting the error-minimized relevant pattern. Figure 2 shows the structure of Nonlinear Sammon Projective Pattern Selection.

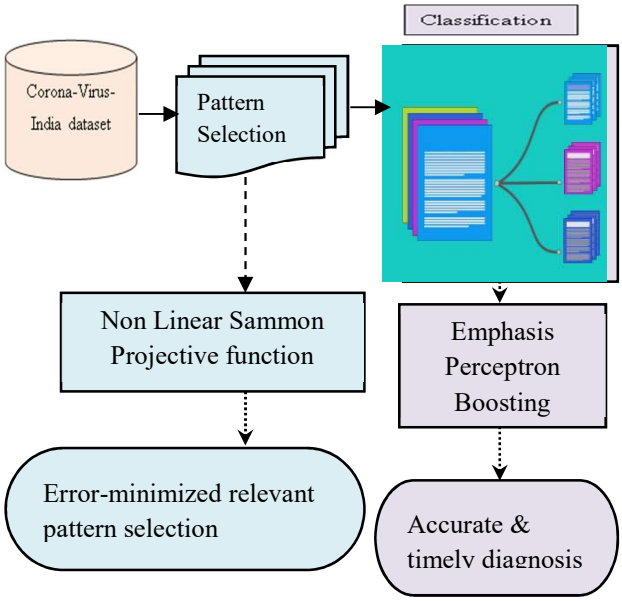

Figure 2: Structure of Nonlinear Sammon Projective Pattern Selection

As shown in the above fig us consider a mapping function ' fun ' that

transforms a pattern 'Q' (as provided in table 1, the latest district level counts) with features'$F = f_1 f_2, .., f_m$', here '$m = 11$' of a 'n' input space (as provided in table 2, the entire details of the dataset) to a pattern 'P' of an 'n' dimensional projected space (i.e., ' $n < m$ ') that is mathematically formulated as given below.

$$P = fun\ (Q) \tag{1}$$

To the above equation (), a criterion 'C' is introduced to optimized relevant pattern selection. Then, the mapping 'f', is determined from among all the transformations ' g ', by satisfying the condition given below.

$$C\ \{fun\ (Q)\} = \max C\ \{g(Q)\} \tag{2}$$

The mappings as given above differ by the functional forms of g and by the criteria they have to optimize (i.e., to preserve the inter-pattern distances). In other words, while performing using the features 'F', the inter-pattern distances should be preserved so that optimized error-minimized relevant pattern selection are made. Then, with the patient files considered as input, obtained from table 3, Nonlinear Sammon Projective Pattern Selection is performed to select relevant patterns for further classification process. Sammon projection being a nonlinear approach maps high-dimensional space, to lowdimensionality space that in turn preserve the inter-point distance structure in high-dimensional space in lower-dimension projection.

The preservation of this inherent inter-point distance structure for each patient files is arrived at by preserving the distances between patterns under projection. Let us consider the inter-pattern distances between pattern 'P_i' and pattern 'P_j' in the input space and in the projected space as '$IS(P_i, P_j)$' and '$PS(P_i, P_j)$' respectively. Here, the pattern in consideration for our work involving the confirmed cases, active cases and recovered cases are modeled.

$$\frac{dC(t)}{dt} = \alpha\ (N) - \mu_1 A(t) \tag{3}$$

From the above equation (3), the confirmed cases at time 't' is arrived at based on the rate at which the susceptible population goes to the confirmed cases '$\alpha\ (N)$' and the susceptible population take onside as active people at time instance ' t ' denoted as '$\mu_1(t)$'. In a similar manner, the active cases at time 't' is measured as given below.

$$\frac{dA(t)}{dt} = \beta\ (N) - \mu_2(t) + \frac{dC(t)}{dt} \tag{4}$$

From the above equation (4), '$dA(t)$' is estimated based on the '$\beta\ (N)$' representing the rate at which the susceptible population goes to active cases from confirmed cases and rate at which confirmed cases '$\frac{dC(t)}{dt}$' are added to active cases ' $\mu_2(t)$ ' respectively. Similarly, the deceased cases '$dD(t)$' are arrived at as given below.

$$\frac{dD(t)}{dt} = \gamma\ (N) \tag{5}$$

From the above equation (5) the deceased cases are obtained based on the rate at which the active cases goes to the deceased cases '$\gamma\ (N)$' respectively. Then, based on the above three patterns in consideration, the patient file data is said to be preserved upon satisfaction of the condition given below using Sammons mapping as given below.

$$E = \frac{1}{\sum_{P_i=1}^{m}\sum_{P_j=1}^{n} IS(P_i,P_j)} \sum_{P_i=1}^{m}\sum_{P_j=1}^{n} \frac{[IS(P_i,P_j) - PS(P_i,P_j)]^2}{IS(P_i,P_j)} \tag{6}$$

From the above equation (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 pseudo code representation of Nonlinear Sammon Projective Pattern Selection is given below.

Algorithm 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 in equation (1) with maximizing criterion as in equation (2) Step 8: For each time instance 't' Step 9: Estimate confirmed cases as in equation (3) Step 10: Estimate active cases as in equation (4) Step 11: Estimate deceased cases as in equation (5) Step 12: **End for** Step 13: Formulate Sammons mapping as in equation (6) Step 14: **Return** (relevant patterns 'RP') Step 15: **End for** Step 16: **End for** Step 17: **End**

As given in the above Nonlinear Sammon Projective Pattern Selection algorithm, the objective remains in selecting optimal and error-minimized relevant pattern. With this objective and three criteria taken into consideration, i.e., confirmed cases, active cases and deceased cases, the patterns observed from infectious disease caused by a newly discovered corona virus can be selected in an effective manner by means of Nonlinear Sammon Projection. By employing sammon's mapping inter-pattern distances are preserved and therefore reduces the error rate involved in selecting the relevant patterns.

3.2Emphasis Perceptron Boosting

Upon successful selection of the optimal and error-minimized relevant pattern, accurate clinical classification of patients with COVID-19 has to be performed. The objective behind the accurate and early classification is that immediate isolation of the COVID-19 patient has to be made so that normal persons will not be infected and with the early classification speed recovery can be ensured. With this objective, Emphasis Perceptron Boosting Classification is applied in our work to categorize the COVID disease patient files with better accuracy and also timely analysis can be performed.

Journal of Theoretical and Applied Information Technology
31st May 2022. Vol.100. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

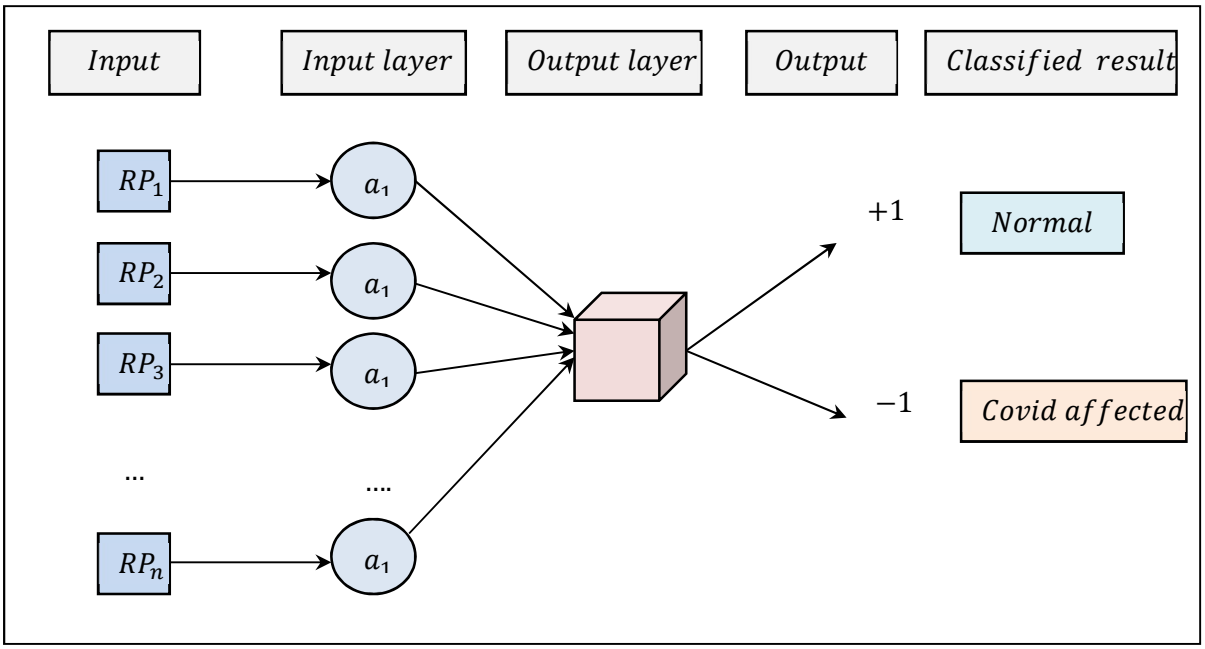

Figure 3: Structure of Emphasis Perceptron Boosting Classification

The Emphasis Boost Classifier in our work merges the weak learner result to form strong classifier output. Here, for patient file classification, perceptron binary classifier is employed. The perceptron binary classifier sees to that whether input denoted by vector of numbers belongs to specific class or not. In other words, it classifies the data points or patterns into two groups diseased or non-diseased. The perceptron binary classifier is a type of linear classifier that performs prediction based on the linear predictor function combining weights with feature vector. This in turn helps to improve the COVID disease prediction with higher accuracy and minimal time consumption. Figure 3 shows the structure of Emphasis Perceptron Boosting Classification model.

The perceptron binary classifier considered as the weak classifier comprises of four parts. They are input values (i.e., relevant patterns 'RP'), weight ('W'), bias ('B') and the activation function ('AF'). As given in the above figure, all the input values or the relevant pattern 'RP', are multiplied with their corresponding weight 'W' as given below.

$$WS = \sum_{i=1}^{n} RP_i W_i = (rp_1 w_1 + rp_2 w_2 + \cdots + rp_n w_n) \quad (7)$$

From the above equation (7), the weighted sum 'WS', is arrived at based on the index 'i' and sigma for summation. As already stated that perceptron are weak learner with linearly separable, there prevails a linear classifier that separates the pattern with zero training error as given below.

$$a^T p + p_0 \begin{cases} > 0, \forall a \text{ in } P_i(\text{ class } [+1]) \\ < 0, \forall a \text{ in } P_i(\text{ class } [-1] \end{cases} \quad (8)$$

From the above equation (8), 'p', 'p_0' symbolizes the parameters that models the linear classifier. Finally, by combining the set of weights 'WS' with feature vectors 'RP', the linear classifier with zero training error is mathematically formulated as given below.

$$fun\,(a; p) = sign\,(a^T, p) \forall i = 1,2,3, \ldots, n \quad (9)$$

Finally, with the objective function to predict the correct label (i.e., diseased patient as diseased and normal patient as normal), for the relevant patterns, the Emphasis Perceptron Boosting algorithm minimize the objective function as given below.

$$obj = -\sum_{i=1}^{n}[b_i fun\,(a;p)]1\,\{b_i \neq fun\,(a;p)\} \quad (10)$$

If the predicted value '$fun\,(a;p)$' and the labels 'b_i' have the same sign then the dot product '$b_i fun\,(a;p)$' would be '> 0'. Then, this refers to that the above defined linear classifier '$fun\,(a;p)$' has predicted the results correctly and accurately for the 'a_i' data-point. The pseudo code representation of Emphasis Perceptron Boosting Classification for early COVID-19 prediction is given below.

Algorithm 2: Emphasis Perceptron Boosting Classification
Input: Dataset 'DS', district level counts 'DLC', Patient-wise data 'PD', patient number 'PID' **Output**: Accurate classified results
Step 1: **Initialize** relevant patterns '$RP = rp_1, rp_2, \dots, rp_n$', Weight '$W = w_1, w_2, \dots, w_n$' **Step 2**: **Begin** **Step 3**: **For** each Dataset 'DS' with district level counts 'DLC' obtained from Patient-wise data 'PD'patient files with patient number 'PID' **Step 4**: Evaluate weighted sum as in equation (7) **Step 5**: Formulate linear classifier as in equation (8) **Step 6**: Formulate weight aggregated linear classifier as in equation (9) **Step 7**: Evaluate objective function as in equation (10) **Step 8**: **Return** classified results 'b_i' **Step 9**: **End for** **Step 10**: **End**

As given in the above Emphasis Perceptron Boosting Classification algorithm, the objective remains in predicting the COVID-19 at an early stage with maximum accuracy. With this objective, a weak classifier based on perceptron function is first modeled. Then, emphasis boosting is formed by means of weighted sum that separates the pattern with zero training error. Finally, by integrating the weight and feature vector objective function is formulated to return the predicted result at an early stage. Followed by which the actual classified results are retrieved in an accurate manner.

4. EXPERIMENTAL SETUP

The objective behind the design of this study is to accurately predict the outcome of a particular patient based on several factors, including but not limited to city in which case is detected, district in which case is detected, current status, contracted from which patient, nationality, type of transmission, source 1, source 2, source 3 etc. As this remains to be a very critical prediction, accuracy is very important. Thus, for the purpose of evaluating the proposed Nonlinear Sammon Projective and Perceptron Boosting (NSP-PB) method for early COVID-19 prediction we considered three evaluation metrics for this study. They are, prediction accuracy, prediction time and error rate with respect to number of patient files.

To perform a fair comparison between the proposed NSP-PB method and existing methods, machine learning forecasting method [1], hybrid feature representation and random forest [2], data are collected from https://www.kaggle.com/imdevskp/covid19-corona-virus-india dataset?select=district_level_latest.csv [20]. Simulations are performed using Python high-level generation purpose programming language. The dataset details are provided in table 1 (district level counts), table 2 (overall csv files) and table 3 (patient-wise data).

Table : Latest district level counts 'DLC'

S. No	Features
1	State name
2	State code
3	Name of the district
4	Number of confirmed cases
5	Number of active cases
6	Number of deceased cases
7	Number of recovered cases
8	Change in confirmed cases
9	Change in active cases
10	Change in deceased cases
11	Change in recovered cases

Table 2: Details of COVID-19 Corona Virus India dataset

S. No	Features	Description
1	Complete	Cumulative count of each day's data from each states
2	District level latest	Latest district level counts
3	Nation level daily	Daily nation level numbers
4	Patients data	Patient-wise data collected from https://api.covid19.org/
5	State level daily	State level daily
6	State level latest	Latest state level
7	Test day wise	Day wise test statistics
8	Test state wise	State wise test statistics

Table 3: Patient-wise data 'PD'

S. No	Features
1	Patient number
2	Patient ID
3	State wise patient ID
4	Date on which case is announced
5	Age
6	Gender
7	City in which case is detected
8	District in which case is detected
9	State code
10	Current status
11	Contracted from which patient
12	Nationality
13	Type of transmission
14	Source 1
15	Source 2
16	Source 3

5. DISCUSSION

In this section the performance analysis of three different parameters namely, error rate, prediction accuracy and prediction time with respect to distinct numbers of patients are provided. To ensure fair comparison all the three parameters were analyzed with similar number of patients using all the three methods, Nonlinear Sammon Projective and Perceptron Boosting (NSP-PB), machine learning forecasting method [1] and hybrid feature representation and random forest [2].

5.1 Performance analysis of error rate

The most paramount feature for analyzing the prediction method is the error rate. This is because of the reason that patient only affected with COVID should be declared as COVID and isolated for further treatment till he has been completely recovered. On the other hand, the patient not affected with COVID should be correctly diagnosed and should not be provided with medication. Hence, the error rate is mathematically formulated as given below.

$$ERate = \sum \frac{PWP_{AC}}{P_{AC}} \quad (11)$$

From the above equation (11), the error rate '$ERate$', is measured based on the patients affected with COVID 'P_{AC}' from an overall population to the patients wrongly predicted as affected with COVID 'PWP_{AC}'. It is measured in terms of percentage (%). Table 6 given below shows the result analysis of prediction accuracy using three different methods, NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2].

Table 4: Error rate analysis using NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2]

Number of patients	Error rate (%)		
	NSP-PB	machine learning forecasting method	hybrid feature representation and random forest
8000	10	15	20
16000	10.35	15.15	20.25
24000	10.55	15.35	20.45
32000	11	15.85	20.75
40000	11.25	16	20.95
48000	11.45	16.15	21
56000	11.85	16.55	21.35
64000	12	16.85	22

72000	12.25	17	22.45
80000	13	17.35	24

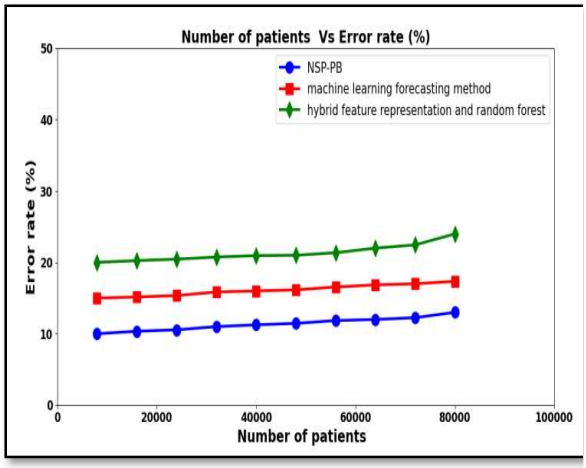

Figure 4: Graphical representation of error rate

First, figure 4 given above shows the error rate involved in predicting the patients into either COVID affected or not affected patients. As illustrated in the above figure, error rate as found to be directly proportion to the number of patients provided as input for simulation. In other words, increasing the number of patients causes an increase in the patients to be analyzed for COVID affected or not affected with COVID. Due to this a small portion or error is said to occur or in other words, a small portion of input are found to be wrongly identified with COVID though they do not have the same and vice versa. However, simulations conducted with 8000 patients and 20 patients actually infected with COVID and wrongly detected as 2 using NSP-PB, 3 and 4 using [1] and [2], the error rate was observed to be 10%, 15% and 20% respectively. With this result it is inferred that the error rate was found to be comparatively lesser than [1] and [2]. The reason behind the improvement was due to the application of Nonlinear Sammon Projective Pattern Selection algorithm. By applying this algorithm, three criteria were taken into consideration using Nonlinear Sammon Projection. With this projection function inter-pattern distances were preserved using NSP-PB that in turn reduced the error rate by 30% compared to [1] and 47% compared to [2].

5.2 Performance analysis of prediction accuracy

The first and foremost parameter of significance is the prediction accuracy or simply the accuracy involved during the early prediction of patients being affected by COVID. Given a dataset consisting of true positive and true negative data points, the accuracy is equal to the ratio of total correct predictions or accurate predicted by the classifier to the total data points or patients involved in the simulation process. Prediction accuracy is hence contemplated as a significant measure which is utilized to estimate the performance of the classification method. Prediction accuracy is calculated as shown in the below equation.

$$PAcc = \sum_{i=1}^{n} \frac{P_{AP}}{P_i} \quad (12)$$

From the above equation (12), prediction accuracy '$PAcc$' is measured based on the patients involved in the simulation process 'P_i' and the patients accurately predicted 'P_{AP}'. It is measured in terms of percentage (%). Table 5 given below shows the result analysis of prediction accuracy using three different methods, NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2].

Table 5: Prediction accuracy analysis using NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2]

Number of patients	Prediction accuracy (%)		
	NSP-PB	machine learning forecasting method	hybrid feature representation and random forest
8000	97.68	97.5	96.68
16000	97.25	93.85	92.15
24000	97	93	92.05
32000	97	93	91.85
40000	96.85	92.25	91
48000	96.35	92.15	91
56000	96	92	90.15
64000	95.85	91.25	90.15
72000	95.15	91.15	90
80000	95	91	89.85

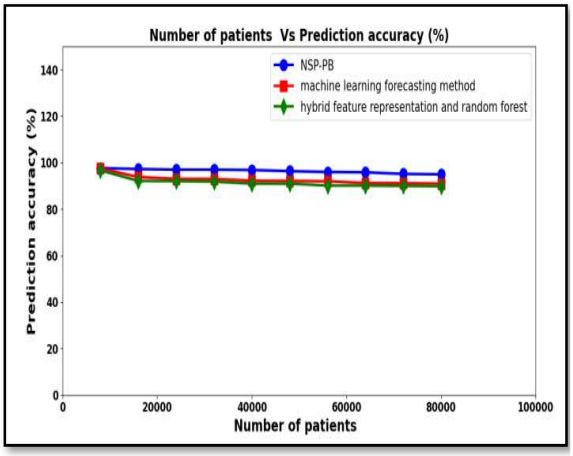

Figure 5: Graphical representation of prediction accuracy

Figure 5 illustrated above shows the graphical representation of prediction accuracy with respect to 80000 numbers of patients obtained from different districts at distinct time intervals. With x axis representing the number of patients and y axis denoting the prediction accuracy, a steep drop in accuracy rate is seen to be observed using the three methods, NSP-PB, [1] and [2]. However, with 8000 numbers of patients involved for simulation and 7815 patients being accurately predicted with actual case (i.e., COVID affected person as so and not affected by COVID as so) using NSP-PB, 7800 using [1] and 7735 using [2], the overall prediction accuracy were found to be 97.68%, 97.5% and 96.68% respectively. With this analytical results it is inferred that the prediction accuracy using NSP-PB upon comparison with [1] and [2]. The reason behind the improvement in prediction accuracy was due to the application of Emphasis Perceptron Boosting Classification algorithm. By applying this algorithm, weak classifier based on perceptron function was initially modeled. Then, by employing emphasis boosting using weighted sum that separates the pattern with zero training error was performed. This in turn resulted in the improvement of prediction accuracy of the respective patients with accuracy recognition of COVID patients using NSP-PB by 4% compared to [1] and 5% compared to [2].

5.3 Performance analysis of prediction time

The second parameter of significance is the prediction time or the time involved in the prediction of whether a patient is affected with COVID or not. This parameter plays an important role in analyzing the overall process as early the prediction is made, earlier the patient can be isolated and hence higher becomes the possibility of recovery. The prediction time is formulated as given below.

$$PTime = \sum_{i=1}^{n} P_i * Time\ [fun\ (a;p)] \quad (13)$$

From the above equation (13), the prediction time '$PTime$' is measured based on the patients involved in simulation 'P_i' and the time involved in classifying the patient affected by COVID or not affected by COVID '$Time\ [fun\ (a;p)]$'. It is measured in terms of milliseconds (ms). Table 6 given below shows the result analysis of prediction time using three different methods, NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2].

Table 6: Prediction time analysis using NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2]

Number of patients	Prediction time (ms)		
	NSP-PB	machine learning forecasting method	hybrid feature representation and random forest
8000	10800	14800	16400
16000	12500	15352	18135
24000	13400	17150	20325
32000	14150	19325	22455
40000	15325	20145	24515
48000	16180	22325	28315
56000	18350	24510	30125
64000	21245	26000	32325
72000	24325	27325	35140
80000	25000	29000	38000

Figure 6 given above illustrates the graphical representation of prediction time using the three distinct methods, NSP-PB, machine learning forecasting method [1] and hybrid feature representation and random forest [2]. With x axis representing the numbers of patients in the range of 8000 to 80000 and the y axis representing the prediction time in terms of milliseconds (ms), with the increase in the numbers of patients provided for simulation as input, the patient data to be analyzed with the deceased, active and recovered cases increases. This in turn results in the increment of the prediction time or the time consumed in predicting the patient either affected with COVID or not. However, simulation results show that with 8000 patients involved in the simulation process, time involved in classifying the patient affected by COVID or not affected by COVID for single patient using NSP-PB was found to be 1.35 ms, 1.85ms and 2.05ms using [1] and [2].

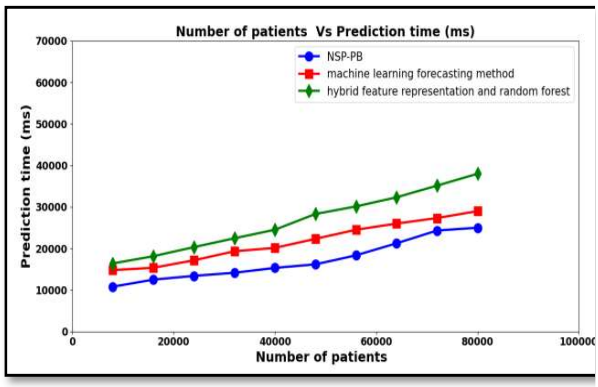

Figure 6: Graphical representation of prediction time

With this result it is inferred that increasing the number of patients causes an increase in the time consumed in analyzing the active/recovered cases and hence causing an increase in the prediction time. However, simulation result showed improvement by using NSP-PB upon comparison with [1] and [2]. The reason behind the improvement was due to the prediction made based on the linear predictor function combining weights with feature vector that in turn reduced the time consumed in predicting using NSP-PB by 21% compared to [1] and 35% compared to [2] respectively.

6. CONCLUSION

This paper proposed Nonlinear Sammon Projective and Perceptron Boosting (NSP-PB) for early COVID-19 prediction in COVID-19 Corona Virus India dataset to overcome the limitation of maximum prediction accuracy and minimum prediction time, error rate. This is achieved by means of two different models, Sugeno Fuzzy Inference model and Model-free Reinforcement Learning model. Prediction error rate is reduced by using Nonlinear Sammon Projective Pattern Selection to analyze the COVID cases for choosing the optimal relevant pattern. Inter-pattern distance is preserved by means of Sammon function. Moreover, the prediction accuracy and time involved in the overall prediction process was said to be reduced with aid of Emphasis Perceptron Boosting Classification. Emphasis Boost Classifier weak learner result are merged to form strong classifier.

The proposed method achieves higher prediction accuracy, consumes very less time and produces minimum error rate as compared to other state-of-the-art methods. The simulation results show that the NSP-PB method accurately and timely predicts the patient data into COVID affected or COVID not affected with minimum error rate. Therefore, the proposed NSP-PB has achieved a higher the prediction accuracy by 9%, lesser prediction time by 28%, and error rate by 39% than the exiting methods.

The proposed NSP-PB is failed to concentrate on evaluating the performance of different parameters, such as classification accuracy, feature selection rate, and memory consumption. But, it fails to present storage system for storing selected patterns and classified data. In future work, proposed technique is further extended to perform various feature selection process to select relevant patterns. Then, various classifier processes is presented to classify data as COVID affected or COVID not affected patients from input database. Also, the classification accuracy, feature selection rate, and memory consumption is measured. Data storage is considered during the selection of relevant patterns.

REFERENCES

[1] R. Sujath, Jyotir Moy Chatterjee and Aboul Ella Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India", *Stochastic Environmental*

Research and Risk Assessment, *Springer*, Vol. 34, 2020, pp. 959 – 972

[2] ChunyanAo, Wenyang Zhou, Lin Gaoa, Benzhi Dong, Liang Yu, "Prediction of antioxidant proteins using hybrid feature representation method and random forest", *Genomics, Elsevier*, Vol. 112, No. 6, 2020, pp. 4666-4674

[3] Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres, "Classification models for heart disease prediction using feature selection and PCA", *Informatics in Medicine Unlocked, Elsevier,* Vol. 19, 2020, pp. 1-11

[4] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam and Gyu Sang Choi, "COVID-19 Future Forecasting using Supervised Machine Learning Models", *IEEE Access*, Vol. 8, 2020, pp. 101489 – 101499

[5] Mohamed Abdel-Basset, Victor Chang, Nada A. Nabeeh, "An intelligent framework using disruptive technologies for COVID-19 analysis", *Technological Forecasting & Social Change, Elsevier*, Vol. 163, 2020, pp. 1-32.

[6] Syed Javeed Pasha, E. Syed Mohamed, "Novel Feature Reduction (NFR) Model With Machine Learning and Data MiningAlgorithms for Effective DiseaseRisk Prediction", *IEEE Access*, Vol. 8, 2020, pp. 184087 - 184108

[7] Sanjay Kumar Sonbhadra, Sonali Agarwal, P. Nagabhushan, "Target specific mining of COVID-19 scholarly articles using one-classapproach", Chaos, Solitons & Fractals, *Elsevier,* Vol. 140, 2020, pp. 1-15.

[8] Srinivas Koppu, Praveen Kumar Reddy Maddikunta, Gautam Srivastava, "Deep learning disease prediction model for use with intelligent robots", *Computers and Electrical Engineering*, *Elsevier,* Vol. 87, 2020, pp. 1-13.

[9] Mohamed Marzouk, Nehal Elshaboury, Amr Abdel-Latif, ShimaaAzab, "Deep learning model for forecasting COVID-19 outbreak in Egypt", *Process Safety and Environmental Protection, Elsevier,* Vol. 153, 2021, pp. 363-375

[10] Haolin Wang, Xuhai Tan, Zhilin Huang, Bo Panc, Jie Tian, "Mining incomplete clinical data for the early assessment of Kawasaki diseasebased on feature clustering and convolutional neural networks", *Artificial Intelligence In Medicine, Elsevier,* Vol. 105, 2020, pp. 1-7.

[11] Olaide N. Oyelade, Absalom E. Ezugwu, "A case-based reasoning framework for early detection and diagnosis of novel coronavirus", *Informatics in Medicine Unlocked, Elsevier,* Vol. 20, 2020, pp. 1-22.

[12] NopharGeifman, Anthony D. Whetton, "A consideration of publication-derived immune-related associationsin Coronavirus and related lung damaging diseases", *Journal of Translational Medicine*, Vol. 18, No: 297, 2020, pp. 1-11.

[13] Farman Ali, Shaker El-Sappagh, S.M. Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran , Kyung-Sup Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion", *Information Fusion, Elsevier*, Vol. 63, 2020, pp. 2-0-222.

[14] Patricia Melin, Julio Cesar Monica, Daniela Sanchez, Oscar Castillo, "A new prediction approach of the COVID-19 virus pandemic behaviorwith a hybrid ensemble modular nonlinear autoregressive neuralnetwork", *Soft Computing, Springer*, 2020, pp. 1-10

[15] Yazeed Zoabi, Shira Deri-Rozov, Noam Shomron, "Machine learning-based prediction of COVID-19 diagnosisbased on symptoms", *Digital Medicine,* Vol. 4, No. 3, 2021, pp. 1-5

[16] ChellammalSurianarayanan, Pethuru Raj Chelliah, "Leveraging Artificial Intelligence (AI) Capabilitiesfor COVID-19 Containment", *New Generation Computing, Springer,* Vol. 39, No. 4, 2021, pp. 717-741

[17] T. Aishwarya, V. Ravi Kumar, "Machine Learning and Deep Learning Approaches to Analyze and Detect COVID-19: A Review", *Computer Science, Springer,* Vol. 2, No. 3, 2021, pp. 1-9.

[18] Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, PrathameshDaphal, Mahesh Warang, NinadMehendale, "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning", *Annals of Data Science, Springer,* Vol. 8, 2020, pp. 1-19

[19] Ameer Sardar Kwekha-Rashid, Heamn N. Abduljabbar, Bilal Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications", *Applied Nanoscience, Springer*, 2021, pp. 1-13.

[20]https://www.kaggle.com/imdevskp/covid19-corona-virus-india dataset?select=district_level_latest.csv
[21] Ahmed S. Salama, Ahmed M. Eassa, "IoT and cloud based blockchain model for covid-19 infection spread control", *Journal of Theoretical and Applied Information Technology,* Vol. 100, No. 1, 2022, pp. 1-14
[22] Mrs. Manju D, Dr. Seetha M, Dr. Sammulal P, "Performance analysis of LSTM based deep learning models for abnormal action prediction in surveillance videos", *Journal of Theoretical and Applied Information Technology,* Vol. 100, No. 1, 2022, pp. 1-11
[23] Sridevi Sakhamuri, Dr. K. Kiran Kumar, "Deep learning and metaheuristic algorithm for effective classification and recognition of paddy leaf diseases", *Journal of Theoretical and Applied Information Technology,* Vol. 100, No. 4, 2022, pp. 1-11
[24] Qianfeng Lin, Jooyoung Son, "An IN-SHIP localization algorithm for close contact identification", *Journal of Theoretical and Applied Information Technology,* Vol. 100, No. 6, 2022, pp. 1-12
[25] V. Laxmi Narasamma, Dr. M. Sreedevi, "Detecting malicious activities on twitter data for sentiment analysis using a novel optimized machine learning approach", *Journal of Theoretical and Applied Information Technology,* Vol. 100, No. 6, 2022, pp. 1-12
[26] JV Vardhan, PC Vemula, G Srinivas, G Chowdary, SB Kumar, "Diagnosis of covid-19 using 3d convolutional neural networks", *Journal of Theoretical and Applied Information Technology,* Vol. 99, No. 24, 2021, pp. 5794-5803