# Hybrid recurrent neural network using custom activation function for COVID-19 short term time series prediction

**Tirupati G,[1*] Krishna Prasad MHM,[2] Srinivasa Rao P[3]**

[1]*Department of Information Technology,GVP College of Engineering for women(A),Visakhapatnam,India*
[2]*Department of Computer Science and Engineering,University College of Engineering(A),Kakinada,India*
[3]*Department of Computer Science and Engineering,MVGR College of Engineering(A),Vizianagaram,India*
*E-mail ID's: - gtr@gvpcew.ac.in, krishnaprasad.mhm@gmail.com, psr.sri@gmail.com*

**KEYWORDS**
Vanishing gradient, Taylor expansion, Functional approximation, Sigmoid, Hyperbolic tangent

**ABSTRACT:**
**Introduction**: The nonlinear behaviour of activation functions is vital in Artificial Neural Networks (ANNs) for exploring the complex relationship between the input and output features. However, these are probably going to encounter vanishing gradient problems due to small gradients that lead to training instability, expensive exponent operations, and slow convergence. **Objectives**: The primary objective of this study is to develop Taylor expansion of the second order to realize the hyperbolic tangent and sigmoid functions. In particular, long short term memory network make extensive use of these functions as well as gating mechanism to control the flow of information and gradients. Both the custom functions can reduce the vanishing gradient issues in recurrent neural networks. **Methods:** Taylor expansion hyperbolic tangent and sigmoid activation functions based parallel heterogeneous Long Short Term Memory Network integrated with Bayesian hyperparameter Optimization is being proposed for coronavirus multi step time series prediction. a Min-Max Normalization is applied, which produces scaled data in the range (0, 1). The normalized dataset is partitioned into training and testing datasets, with 80% and 20%, respectively. Furthermore, both train and test datasets are prepared as input and target series using a window size of 5-7.The further proposed model is tuned with key hyperparameters such as the number of neurons, learning rate, dropout, and type of optimizer. The remaining model parameters are epochs, batch size, and loss, which are 200, 32, and mean square error, respectively. **Results**: The proposed model efficacy is evaluated on coronavirus daily cumulative cases, cumulative deaths, daily new cases, and total recovery cases in India. The Analysis reveals that the current model achieves remarkable performance in terms of Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of determination ($R^2$ Score) when compared to existing models. **Conclusions**: The study reveals that the proposed framework with the Taylor approximation activation function produces more consistency in prediction than the default activation functions, including Tanh and sigmoid. In spite of that, gradients of Taylor Tanh and sigmoid activation function traits indicate a decline in the possibility of vanishing issue.

## 1. Introduction

Nonlinear behaviour of activation functions is vital in Artificial Neural Networks (ANNs) for exploring the complex relationship between the input and output features. Machine learning, entrusted with the use of Artificial Neural Networks for regression [14] and multi labelling [3] has been adapted in recent times. These comprise forecasting [1], image analysis [23] machine translation [16] and many more. A neuron is responsible for two major operations, including product-accumulation and nonlinear function. The activation function incorporates nonlinearity to make Neural Networks are successful in versatile applications. The popular activation functions include Hyperbolic Tangent (Tanh), Sigmoid, Rectified Linear Unit (ReLU) family and Softmax. With the bounded activities, the prominent activation functions are Sigmoid and Tanh is becoming accustomed in most of the tasks [28]. These are significant and extensively used as gating operations in Recurrent Neural Networks (RNNs). ReLU family is prominent for unbounded activation functions and softmax is wide spread in the output layer including multi label

categorization and attention process. Therefore, Tanh and Sigmoid activation functions are used as activation and recurrent activation for regression tasks. However, these are probably going to encounter vanishing gradient problems due to small gradients that lead to training instability, expensive exponent operations and slow convergence [17].
During training, derivative weights approach zero that causes vanishing gradient problem. As a result, it is difficult to update the weights using Back propagation algorithm. Another challenge is the cost of hardware to implement the exponent operations [40]. Hence, enhancing derivatives of these functions will indefinitely reduce vanishing gradient difficulties and increase the efficiency. To alleviate such problems, several methods have been communicated. These include variants for sigmoid [34], Tanh [7,10], cost reduction on Field Programmable Gated Arrays [45], Normalization Techniques [26] and gradient clipping techniques [43]. The above said methods mostly concentrate on declining the exponent operations and avoiding weights approaching zero instead of enhancing the gradients. Moreover, the performance is changing in different applications in the same domain. These problems can be addressed by popular methods including appropriate activation functions and gated architectures. Thus, a second order Taylor Tanh and sigmoid functions based parallel heterogeneous Long short term Memory (LSTM) architecture is being proposed for multi-step time series prediction.
The work has been structured as follows: Part 2 designates the current work associated with the domain of vanishing gradient. In Part 3, we describe the dataset in this work followed by the methods and experimental design in Part 4 and results and discussions of study presented in Part 5. Finally, in part 6 conclusions and future research directions have been forwarded.

## 2. Literature Survey

In ANN, the vanishing gradient issue can be addressed by Residual Neural Networks (ResNets) that enable the derivative flow and reduce training difficulties. In the same way, a novel approach [26] enhances the ResNets by incorporating norm(magnitude) preservation mechanism. The authors provide insights into how maintaining norm consistency can address challenges related to vanishing gradients and improve overall network performance. However, the addition of norm preservation mechanisms could potentially increase the computational overhead and the approach might face challenges in scaling to more complex architectures or larger datasets. An innovative approach [19] adjusts the gradient flow to ensure more stable and effective learning by replacing the original derivative with an artificial derivative. This process effectively solves the vanishing gradient problem for both ReLU and sigmoid functions with minimal processing costs. In spite of that, artificial derivative is a heuristic and a unique design is required for each activation function. To decline the exponent operation, recent work [39] promotes implementing low complexity hardware development using Probability based Sigmoid Function (P-SFA) approximation. P-SFA is efficient in terms of power, speed and recognition performance on certain datasets. However, method has not discussed about vanishing gradient difficulties.

The functional approximation has wide spread in recent times in order to reduce cost and vanishing derivative issues. The author [10] outlined that high-order Sobolev norms are used to approximate Tanh activation functions in neural networks. The findings indicate that Tanh neural networks with only two hidden layers can approximate functions at comparable or higher rates than considerably deeper ReLU neural networks. It has been addressed only as an alternate to ReLU function in the prediction.
The Vanishing gradients issue is addressed by clipping the gradients, the author [43] develops a frame work to explore how gradient clipping impacts the network training. The limitations of the framework are Implementing adaptive gradient clipping thresholds can be complex and may require additional computational overhead to determine the optimal clipping values dynamically. The author [17] present theoretical insights into why typical RNNs struggle with vanishing gradients and how their proposed LSTM architecture addresses these concerns. They address the function of gating mechanisms in maintaining gradients and increasing learning outcomes.
A survey[21] discusses the applications of various activation functions across different types of neural networks and tasks. It provides insights into how different functions impact network performance and training dynamics. In [13] review a wide range of activation functions and their role in conducting a

benchmark study. They analyze metrics such as convergence speed, training stability, and overall performance on standard benchmarks.

The author [43] described an approximation of sigmoid for neural networks on Field Programmable Gated Arrays (FPGA). The proposed method showed better results than the original sigmoid. Although it has simplicity of implementation, it does not address the gradient problems. The Progress of neural network training time has been improved in the study [34] on approximate ReLU, Tanh, and sigmoid on three networks: MNIST classifier, MNIST auto encoder, and CharRNN. These ranged approximations produced low training times but are not suitable for other networks. The study [7] explores approximating hyperbolic tangent using Catmull-Rom spline interpolation. The achieved results show that a smaller logic area is required than the original Tanh. Although it declines the cost, it does not discuss the gradient issue. On the other hand, a study [8] used sigmoid approximation using Taylor expansion in multilayer networks for hepatitis disease diagnosis. This implementation has utilized three intervals to approximate the sigmoid. It has achieved equal accuracy with the original sigmoid. The limitations are the complexity and overhead of hardware implementation. The study also focused on functional approximations with respect to softmax alternatives. On the same line, a work [40] investigated periodic alternatives for softmax with attention for escaping the gradient problem. It has been suggested that the periodic activation function better than the original softmax. Another study [4] described how soft-margin Taylor has been put forward as an alternative to softmax. The analysis shows that Taylor expansion up to two terms produces higher accuracy on three datasets (MNIST, CIFIR10, and CIFIR100) than Taylor expansion up to ten terms. The work [35] reports that toluene gas concentrations are estimated using neural network (NN) architectures with Taylor series expansions of the sigmoid activation function, based on transient sensor responses. The Taylor expansion up to nine terms is used to realize the sigmoid function. The part of the literature review is presented in Table 1.

Table 1. A section of literature review pertains to our work

| Reference No. | Model | Description |
|---|---|---|
| [2] | Low-error digital hardware implementation of artificial neuron activation functions and their derivative | In this, a piecewise linear method used to approximate the sigmoid and hyperbolic tangent. It requires few cycles to perform approximations and can be simply pipelined than iterative approach. |
| [9] | Fast and accurate deep network learning by Exponential Linear Units (ELUs) | ELUs alleviate the vanishing gradient problem via the identity for positive values and these have improved learning characteristics compared to the units with other ReLu family. |
| [20] | hyperbolic tangent implementation in hardware: polynomial modeling | In this modeling, uses fractional exponential part and it is faster than CORDIC but slower than the piecewise linear solution. |
| [30] | Function approximation using Look Up Table (LUT) | The LUT approach keeps the sampled values of sigmoid function in RAM or ROM in order to reduce the cost of the operation. |
| [31] | Searching for activation functions | proposed to leverage automatic search techniques to discover new activation functions |
| [32] | Approximation of the sigmoid function and its derivative using a minimax approach | Minmax technique is to reduce the error and simplicity of implementation. Sigmoid and its derivative |

| | | |
|---|---|---|
| | | approximations were implemented in VHDL and synthesized to an FPGA. |
| [38] | Hardware implementation of neural network with Sigmoidal activation functions using Coordinate Rotation Digital Computer(CORDIC) | It is an iterative approach that uses multiple iterations to generate complex operations into a single one. |
| [44] | An Analog continuous valued number system (CVNS)-Based Sigmoid Neuron for Precise Neurochips. | It is implemented based on piecewise linear approximation in the analog domain and proposed sigmoid achieves high accuracy that cannot be attained using analog network. |

## 3. Dataset

To assess the performance of the proposed model for multi-step time series forecasting, we considered six different sizes of coronavirus univariate datasets have been taken from the Oxford Martin Programme on Global Development at Oxford Martin School. For this study, one dataset [29] between February 24, 2020 to May 20, 2020 for India consists of daily cumulative confirmation cases, cumulative deaths, and new cases. The other dataset [36] between January 30, 2020 to August 11, 2021 for India comprises daily cumulative confirmation cases, cumulative deaths, and cumulative recovery cases

## 4. Methods

This section explores the various stages, including data preprocessing, proposed model, and experimental design.

### 4.1 Preprocessing

In order to effectively converge the parameters of the Machine Learning model, the original values are normalized from a wide to a narrow range. For this, Min-Max Normalization is applied as shown in Equation (1), which produces scaled data in the range (0, 1). The normalized dataset is partitioned into training and testing datasets, with 80% and 20%, respectively. Furthermore, both train and test datasets are prepared as input and target series using a window size of 5-7.

$$y' = \frac{y - \min_x}{\max_x - \min_x}(new\_max - new\_min) + new\_min \qquad (1)$$

$y'$ = Normalized value of y
y = Observed value of x
$\min_x$ = Minimal of x
$\max_x$ = Maximal of x
new_max = Maximal of Normalized data
new_ min = Minimal of Normalized data

### 4.2 Proposed Methodology

This section discusses about the proposed activation functions, model, and Bayesian hyperparameter optimization.

### 4.2.1 Taylor activation functions

In neural networks, hyperbolic tangent and sigmoid functions are extensively used in LSTM Networks for controlling the flow of gradients. Hyperbolic Tangent is one of the popular activation functions that takes input as real numbers and produce the output in the range from -1 to 1. It is described in Equation [2]

$$f(x) = \text{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad (2)$$

The Sigmoid activation function takes the real numbers as an input and gives an output in the range of 0 to 1. The expression is given in Equation (3).

$$g(x) = \text{sigmoid}(x) = \frac{1}{(1 + e^{-x})} \qquad (3)$$

Both Tanh and sigmoid are vulnerable to vanishing gradient problems [19] that result in non convergence and reduce efficiency or failure in training.

Taylor Hyperbolic Tangent and Sigmoid: Taylor series approximating the function as a sum of infinite terms defined in terms of the function's derivatives at that point. The author [38] developed a Taylor realization of softmax up to two terms for $e^{x}$ as $1+x+0.5\,x^{2}$. The taylor expansion of $e^{-x}$ and $e^{x}$ is given in Equation (4,5 )

$$e^{-x} = 1 - x + \frac{x^{2}}{2!} - \frac{x^{3}}{3!} + \cdots + \frac{(-1)^{n}x^{n}}{n!} \qquad (4)$$

$$e^{x} = 1 + x + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \cdots + \frac{x^{n}}{n!} \qquad (5)$$

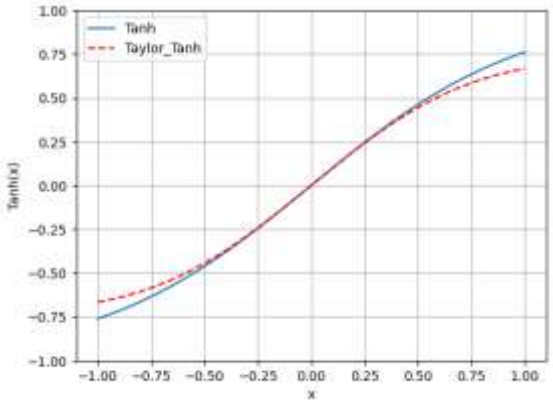


Figure 1. Hyperbolic Tangent (Tanh) and Taylor Hyperbolic Tangent Functions

In this work, n is chosen as 2 for the realization of both hyperbolic tangent and sigmoid functions. The approximation of Tanh activation function using Taylor expansion is shown in Figure 1. The gradients of these functions are primarily responsible for determining the weights using back propagation while training. The derivatives of sigmoid and Tanh functions that are used in back propagation are given in Equation (6, 7). Taylor Tanh and sigmoid functions are contributing enhanced gradients as compared to original Tanh and sigmoid which is shown in Figure 2. As a result, a neural network model with these functions might improve the learning efficiency and reduce the vanishing gradient problem.

$$g'(x) = \text{sigmoid}(x)\big(1 - \text{sigmoid}(x)\big) \qquad (6)$$

$$f'(x) = 1 - tanh^{2}(x) \qquad (7)$$

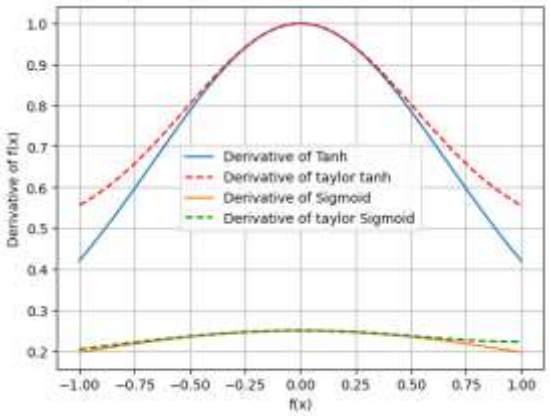


Figure 2. Gradients of Taylor based functions and actual functions

### 4.2.2 Proposed Model

RNN variants are being used successfully in many domains of time series dataset prediction [5,12]. LSTM networks can overcome the limitations of vanilla RNN such as diminishing gradient and exploding problem for long time series. In particular, long short term memory network makes extensive use of Tanh and sigmoid functions as well as gating mechanism to control the flow of information and gradients. However, Vanilla LSTM and bidirectional networks are limited in their performance due to the complex relationships they extract and the large forecasting window size. Moreover, Bidirectional networks have two similar networks that are in parallel; one network takes the input in sequential and the other in an anti-chronological direction. As a result, bidirectional network may not retain the complex relationship in some of the datasets. Therefore, recent studies show that developing hybrid RNN variant Models can produce low prediction errors and high efficiency even in disparate domains and datasets. Hence, we propose a model that consists of two different LSTM networks arranged in parallel fashion, as shown in Figure 3 with Taylor Tanh and Sigmoid activation, recurrent activation, respectively. Each LSTM can uniquely retain the relationship, and then regularization is applied to each LSTM Network to avoid the model over fitting by implementing the dropout approach. Finally, individual LSTM outputs are concatenated, and a dense network is used to predict the output. The structure of the LSTM Cell [18] is shown in Figure 4, and the internal operations of the forget, input, and output gates of the LSTM are performed according to Equation (8) - (13).

$$f_t = \sigma(W_f x_t + W_f h_{t-1} + b_f) \quad (8)$$

$$i_t = \sigma(W_i x_t + W_i h_{t-1} + b_i) \quad (9)$$

$$\tilde{C}_t = \tanh(W_c x_t + W_c h_{t-1} + b_c) \quad (10)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (11)$$

$$O_t = \sigma(W_O x_t + W_O h_{t-1} + b_O) \quad (12)$$

$$h_t = O_t * \tanh(C_t) \quad (13)$$

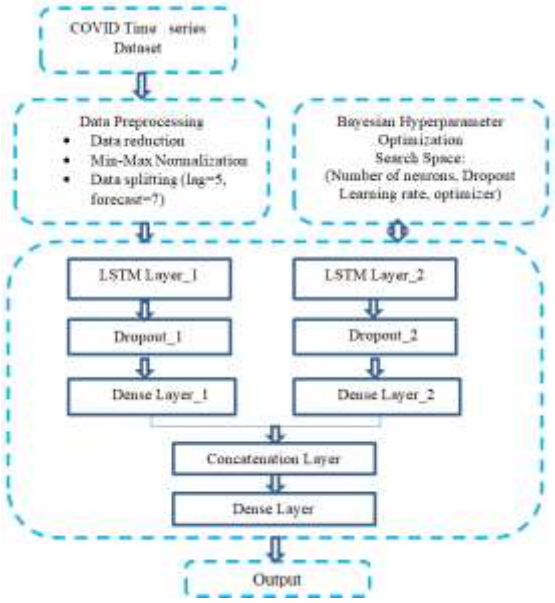


Figure 3. Architecture of proposed model

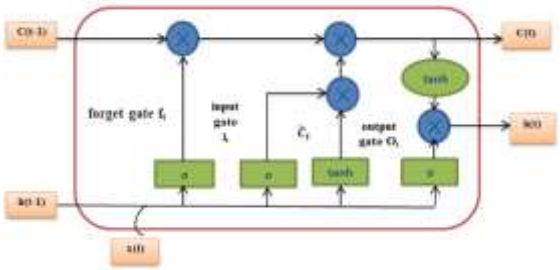


Figure 4. Structure of LSTM Cell

Bayesian hyperparameter Optimization (BO) probabilistic model, which selects current model parameters based on past evaluations. BO is an iterative procedure that is usually used to detect near optimal hyperparameter combinations in just a few iterations [11]. The limitations of both Grid search and Random search are that each evaluation is independent of the previous evaluation. Thus, they spent more time evaluation inadequately performing regions of search space. Bayesian Hyperparameter optimization is more effective than grid search, random search and even manual optimization by domain experts [33]. In Genetic Algorithm, randomly initialized values for genetic operations like crossover, selection, and mutation often do not find the optimal parameters [24]. The main limitation of particle swam optimization is that it requires correct population initialization; otherwise, it may get local optimum instead of global. Bayesian hyperparameter optimization based on Gaussian produces high accuracy and reduces run time when compared to grid and manual search [42,46]. It is preferable when the evaluation of each hyperparameter configuration is time consuming and requires limited resources. The proposed model is tuned using an Automatic Bayesian optimization [28], in contrast to other hyperparameter optimizations [7, 27]. The current model is tuned with respect to search space, as shown in Table 2.

Table 2. Search space for proposed Model

| Name of the parameter | Specifications |
|---|---|
| LSTM Network (Number of neurons) | 4-256 |
| Dropout | 0.1-0.5 |
| Learning rate | 0.001-0.1 |
| Optimizer | Adam, RMSProp |

### 4.2.3 Experimental Design

The proposed model with Taylor Tanh and sigmoid activations is executed in the Google Colab Intel(R), Xenon(R) CPU@2.20GHz in the Python 3.10.12 environment, and other libraries including Tensorflow- 2.15.0, Keras-2.15, Bayesian optimization-1.4.3, automatic ARIMA library pmdarima-2.0.4 and AutoML framework tpot-0.12.2. In the first step, COVID-19 dataset is loaded and then filtered for attributes such as Daily Total Cumulative Confirmation Cases, Cumulative Deaths, new cases and recovery cases for a given duration. The filtered dataset has been processed for NAN values. The resultant dataset is divided into train and test, 80% and 20%, respectively. To train the proposed model, both train and test datasets are split as input and targets. In this splitting process, the lag value is considered to be 5 and the future forecast is 7. The further proposed model is tuned with key hyper parameters such as number of neurons, learning rate, dropout, and type of optimizer. The remaining model parameters are epochs, batch size and loss, 200, 32, and mean square error, respectively. The proposed model with Taylor Tanh and sigmoid activations is tuned using train dataset in 10 iterations, and then BO suggests best hyperparameters. During the evaluation, model uses those best parameters to predict the test dataset. Other models, BLSTM and LSTM are tuned with same search space by Bayesian optimizer and other specifications have default Tanh and sigmoid. The ARIMA Model is developed using pmdarima library and p, d, q values are selected using the auto_arima method for all datasets. An AutoML Model, Tree based Pipeline Optimization (TPOT) repressor selects the best parameters for a given population and generations. The population and generation value are five considered for the TPOT Regressor model fitting.

### 4.3. Evaluation criteria

The proposed model and other models performance is quantified in the form of the Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), R-Squared ($R^2$) Score, and Root Mean Square Error (RMSE).

Mean Absolute Percentage Error (MAPE): MAPE is a popular metric that measures the accuracy of a regression model in terms of relative error[21,41]. It computes the difference between the actual data and the model's forecast

based on test data, then divides that difference by the actual data. A lower MAPE reflects a model prediction that is reasonably close to the observed value. The best MAPE value is zero, and the worst value is infinity. MAPE is characterized as follows;

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \qquad (14)$$

$y_i$, $\widehat{y_i}$ and n is the actual value, prediction value, and size of the data, respectively

Mean Absolute Error (MAE): Another metric is used to assess the regression model, along with MAPE. Sometimes MAE is very different depending on whether if it is predicted or the actual value that is largest. The best score is zero and worst is infinity. MAE is calculated as follows

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (15)$$

R-Squared ($R^2$) Score: $R^2$ score is known as the coefficient of determination and conveys how closely data members fit the curve [22]. The $R^2$ score illustrates how the data values are dispersed over the regression curve. Typically, it lies between -∞ and 1. A high $R^2$ score suggests that the prediction model is reliable. Using Eq. (16), the $R^2$ score can be calculated as follows:

$$R^2 score = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \qquad (16)$$

Root Mean Square Error (RMSE): It is referred to as the residual, and it calculates the prediction error based on the separation between best fit data and actual data. The formula for RMSE is given Eq. (17).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (17)$$

## 5. Results & Discussions

The proposed model with Taylor Tanh and sigmoid activation is evaluated on univariate datasets such as Coronavirus Cumulative Confirmation (CCC) cases (dataset 1), daily New Cases (NC)(dataset 2),Cumulative Deaths(CD) (dataset 3)[26],Coronavirus Cumulative Confirmation(dataset 4),Cumulative Deaths(CD) (dataset 5),and Cumulative Recovery Cases(CRC) (dataset 6) [30] in India.

### 5.1 Coronavirus Cumulative Confirmation (CCC) Cases:

Coronavirus Cumulative Confirmation (CCC) cases in India dataset has been taken from 24th Feb, 2020 to 20th May, 2020. It has been preprocessed and the proposed model is tuned with train data using BO. The best set of hyperparameters has been chosen at the end of 10 evaluations. The proposed model uses the best parameters while predicting 1-week ahead of the CCC prediction. The model produces MAPE, MAE, RMSE, and $R^2$ score values of 1.4570, 0.0109, 0.0130, and 0.9703, respectively. Table 3 shows the performance metrics of the proposed and existing methods for one week ahead CCC prediction. It appears that the present model is superior to model with Tanh and sigmoid activation as well as other existing models. The obtained results of the current model indicate a significant development in prediction as compared to the Deep Learning model with Grey Wolf Optimizer (GWO) tuning [29].The proposed model with Taylor Tanh and Sigmoid 1-week ahead CCC prediction as shown in Figure 5.The model with default Tanh and sigmoid activations shows better than existing recurrent models like BLSTM and LSTM with default Tanh and sigmoid activations, TPOT regressor and ARIMA model. Figure 6 represents low MAPE with respect to Taylor expansion activation functions on the given datasets. The proposed and existing model for one-week ahead CCC prediction is given in Figure 7.

Table 3. One week ahead CCC prediction Performance Metrics of proposed and other models

| Model | MAPE (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|
| ARIMA (021) | 17.0205 | 0.1273 | 0.1541 | 0.2986 |
| TPOT Regressor | 12.4757 | 0.0931 | 0.1769 | -4.7532 |
| BLSTM-BO | 5.4004 | 0.0397 | 0.0431 | 0.6427 |
| LSTM-BO | 2.3250 | 0.0163 | 0.0197 | 0.9220 |
| Pro.Model-BO | 2.0671 | 0.0152 | 0.0167 | 0.9516 |
| Pro.Model(Taylor Tanh & sigmoid)-BO | 1.4570 | 0.0109 | 0.0130 | 0.9703 |

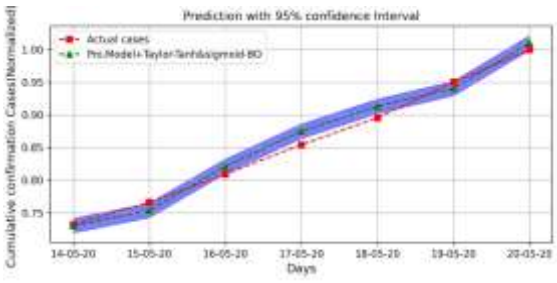


Figure 5. One week ahead CCC prediction of proposed model with Taylor Tanh and Sigmoid

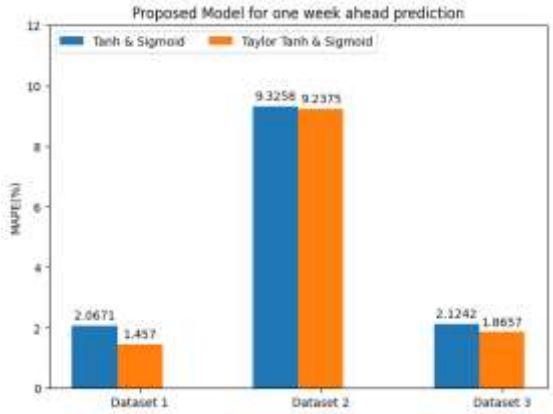


Figure 6. MAPE of the proposed model with Taylor approximation and original activation functions.

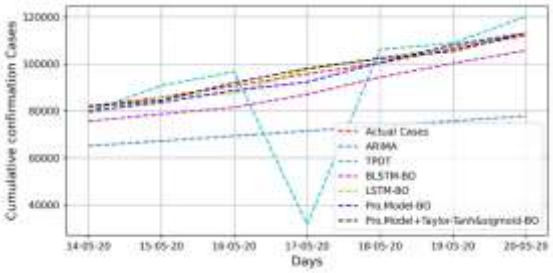


Figure 7. One week ahead CCC prediction of proposed model and other models.

**5.2 Coronavirus daily New Cases (NC):**

For the New Cases (NC) in India dataset has been taken from 24th Feb, 2020 to 20th May, 2020. It has been preprocessed and proposed model with Taylor Tanh and sigmoid is tuned with train data using BO. The best set of configuration parameters is chosen for the subsequent prediction. For the NC, the current model obtains a MAPE of 9.2375, a MAE of 0.0646, a RMSE of 0.0817, and the $R^2$ score of 0. 1756.The results resembles that Taylor Tanh and sigmoid perform slightly better than Tanh and sigmoid functions in the proposed model. Table 4 represents the performance metrics of the proposed and existing models on one week ahead NC prediction. The NC prediction using the current model registers low MAPE as compared to the Deep Learning model with Grey Wolf Optimizer (GWO) tuning [29]. The proposed model with Taylor Tanh and Sigmoid one week ahead NC prediction as shown in Figure 8. The proposed and existing models one week ahead NC prediction is given in Figure 9.

Table 4. One week ahead NC prediction Metrics of Proposed and other models

| Model | MAPE (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|
| ARIMA (013) | 33.4896 | 0.2358 | 0.2738 | -2.5984 |
| TPOT Regressor | 28.1645 | 0.2048 | 0.2729 | -5.4537 |
| BLSTM-BO | 9.4968 | 0.0657 | 0.0831 | 0.1708 |
| LSTM-BO | 10.3363 | 0.0703 | 0.0860 | 0.1156 |
| Pro.Model-BO | 9.3258 | 0.0648 | 0.0826 | 0.1737 |
| Pro.Model(Taylor Tanh & Sigmoid)-BO | 9.2375 | 0.0646 | 0.0817 | 0.1756 |

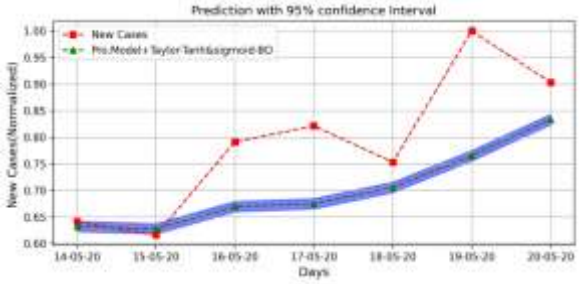


Figure 8. One week ahead NC prediction of proposed model with Taylor Tanh and Sigmoid

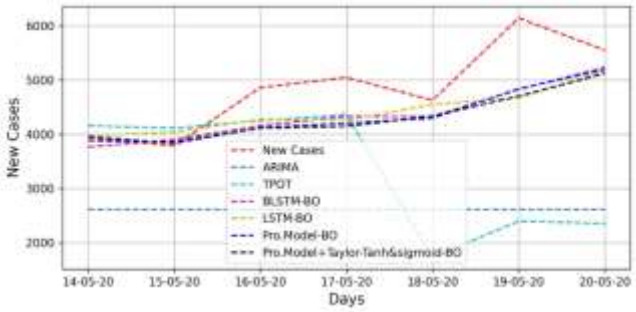


Figure 9. One week ahead NC prediction of proposed model and other models

**5.3 Cumulative Deaths (CD)**

Another dataset is the cumulative Deaths in India, it has been considered from 24th Feb, 2020 to 20th May, 2020.The Proposed model is tuned with a Bayesian optimizer and then best configuration parameters are used for model prediction. The Model performance is measured in terms of MAPE, MAE, RMSE, and $R^2$ score of 1.8657, 0.0144, 0.0180, and 0.9315, respectively. Table 5 shows the low values of MAPE, MAE, RMSE and the high $R^2$ score of the current model with Taylor Tanh and sigmoid activation in comparison to other models. The proposed model obtains lower MAPE than LSTM with Google trends [29]. The proposed model with Taylor Tanh and Sigmoid one week ahead CD prediction as shown in Figure 10. The proposed and existing models one week ahead CD prediction is shown in Figure 11.

Table 5. One week ahead CD prediction Metrics of Proposed model and other models

| Model | MAPE (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|
| ARIMA (220) | 10.5459 | 0.0804 | 0.0931 | 0.7137 |
| TPOT Regressor | 21.0852 | 0.1687 | 0.2608 | -13.002 |
| BLSTM-BO | 2.8992 | 0.0223 | 0.0260 | 0.8493 |
| LSTM-BO | 3.1877 | 0.0237 | 0.0289 | 0.8072 |
| Pro.Model-BO | 2.1242 | 0.0169 | 0.0212 | 0.9156 |
| Pro.Model(Taylor Tanh & Sigmoid)-BO | 1.8657 | 0.0144 | 0.0180 | 0.9315 |

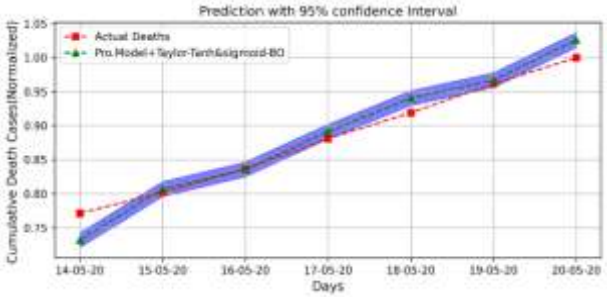


Figure 10. One week ahead CD prediction of proposed model with Taylor Tanh and Sigmoid

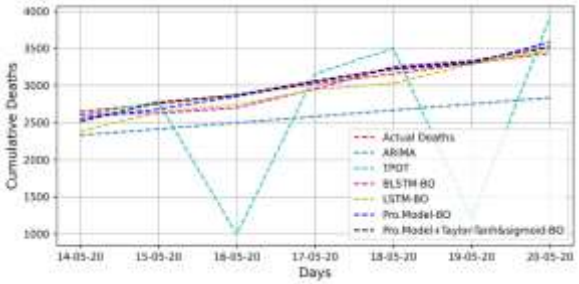


Figure 11. One week ahead TCD prediction of proposed model and other models

**5.4 Coronavirus Cumulative Confirmation (CCC) Cases**

The current model with Taylor Tanh and sigmoid using Bayesian Optimizer is applied to daily Total Cumulative Confirmation (TCC) Cases dataset, which has been considered from 30th January, 2020 to 11th august 2021. The normalized dataset is divided into train, and test data, with 80% and 20%, respectively. Furthermore, both train and test datasets are prepared as input and target series using a window size of 5-7. The Proposed model on this dataset produces a MAPE of 2.0655, a MAE of 0.0180, a RMSE of 0.0224, and an $R^2$ score of 0.9555. Table 6 indicates that proposed model performs better in terms of MAPE, MAE, RMSE and $R^2$ score than other models. Figure 12 shows the superiority of proposed model on three types of datasets. The current model with Taylor approximation of Tanh and sigmoid contributes to improvement in prediction efficiency. Figure 13(a) reveals the one week ahead prediction by the present model and other models.

Table 6. One week ahead CCC prediction Metrics of Proposed model and other models

| Model | MAPE (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|
| ARIMA (020) | 20.4712 | 0.1963 | 0.2740 | -3.1329 |
| TPOT Regressor | 2.4388 | 0.0192 | 0.0261 | 0.9330 |
| BLSTM-BO | 4.7269 | 0.0433 | 0.0482 | 0.7894 |
| LSTM-BO | 2.8126 | 0.0257 | 0.0348 | 0.8851 |
| Pro.Model-BO | 2.7505 | 0.0243 | 0.0297 | 0.9143 |
| Pro.Model(Taylor-Tanh & sigmoid)-BO | 2.0655 | 0.0180 | 0.0224 | 0.9555 |

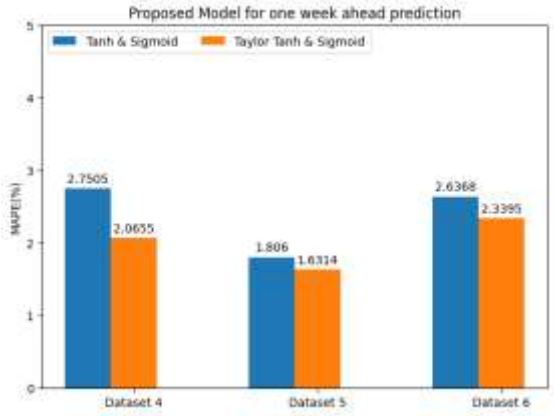


Figure 12. MPAE of the proposed model with Taylor approximation and original activation functions.

**5.5 Cumulative Deaths (CD)**

Another dataset is Total Cumulative Deaths (CD) in India, which has been taken from 30th January, 2020 to 11th august 2021.The proposed model is tuned with a Bayesian optimizer and then best configuration parameters are used for model prediction on cumulative deaths. The Model performance is measured in terms of MAPE, MAE, RMSE, and $R^2$ score of 1.6314, 0.0123, 0.0159, and 0.9882, respectively. Table 7 shows one week ahead deaths prediction metrics for proposed and other existing models. The current model produces low MAPE, MAE, RMSE and the high $R^2$ score. The proposed and existing models one week ahead CD prediction is shown in Figure 13(b)

Table 7. One week ahead CD prediction Metrics of Proposed model and other models

| Model | MAPE (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|
| ARIMA (021) | 14.9573 | 0.1260 | 0.1389 | 0.3798 |
| TPOT Regressor | 52.6570 | 0.4546 | 0.4794 | -9.0208 |
| BLSTM-BO | 8.2617 | 0.0698 | 0.0725 | 0.7675 |
| LSTM-BO | 14.7701 | 0.1285 | 0.1371 | 0.1824 |
| Pro.Model-BO | 1.8060 | 0.0137 | 0.0175 | 0.9858 |
| Pro.Model(Taylor-Tanh & Sigmoid)-BO | 1.6314 | 0.0123 | 0.0159 | 0.9882 |

**5.6 Cumulative Recovery cases (CRC):**

Another coronavirus dataset in India, Cumulative Recovery (CR) cases, has been considered from 30th January, 2020 to 11th august 2021. A total of 560 days of CR cases were obtained and then preprocessed with Min-Max normalization. A processed train data has been utilized in the proposed model tuning, and the selected best parameters are used for model Cumulative Recovery cases for future prediction. Table 8 indicates the performance metrics of current and existing models. It resembles the proposed model with Taylor expansion function, which enhances the prediction accuracy. In addition to Taylor approximations, current model with default Tanh and sigmoid activations exhibits low prediction error. Figure 13(c) shows the one week ahead CR prediction of proposed and existing models.

Table 8. One week ahead CRC prediction Metrics of Proposed model and other models

| Model | MAPE (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|
| ARIMA (021) | 30.6506 | 0.2954 | 0.4150 | -6.2878 |
| TPOT Regressor | 3.1081 | 0.0231 | 0.0331 | 0.9237 |
| BLSTM-BO | 3.3980 | 0.0273 | 0.0339 | 0.9188 |
| LSTM-BO | 5.4033 | 0.0477 | 0.0506 | 0.8377 |
| Pro.Model-BO | 2.6368 | 0.0204 | 0.0283 | 0.9426 |
| Pro.Model(Taylor-Tanh & sigmoid)-BO | 2.3395 | 0.0189 | 0.0247 | 0.9565 |

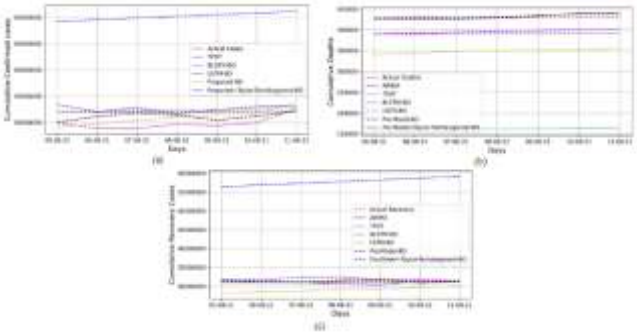

Figure 13. a) One week ahead CCC prediction b) CD prediction c) CRC prediction

## 6. Conclusions & Future scope

In this work, Taylor expansion Tanh and sigmoid activation function based new framework was developed and evaluated on six different sizes of coronavirus time series datasets. The results reveal that the proposed framework with Taylor approximation activation function produces more consistency in prediction than the default activation functions including Tanh and sigmoid. In spite of that, gradients of Taylor Tanh and sigmoid activation function traits indicate decline the possibility of vanishing issues. The current model exhibits high performance on CCC, CD and CRC. Although, NC has nonlinear behavior, present model performs competitive than others. It was observed that proposed model with default Tanh and sigmoid shows low MAPE, MAE, RMSE and High $R^2$ score in contrast to BLSTM and LSTM with default activations. The statistical ARIMA model produces high MAPE, MAE, RMSE and low $R^2$ score on all types of datasets. Auto ML TPOT Regressor gives competitive performance on CCC and CD but poor metrics on other datasets, Current model can effectively be utilized for short term prediction where little past information is known.

The proposed framework directs an initiative towards the development of alternatives to activation functions in Deep Learning. On the other hand, the study gives an initiative for enhancing the gradients of activation functions to avoid learning and declining the exponent operations. However, the scope of experiments might be limited. Further validation across a broader range of datasets and tasks is necessary to fully understand the robustness and generalizability of the proposed method. Concurrently, BO can combine with other popular Hyper-Parameter Optimization (HPO) algorithms in order to enable parallelization

**References**

[1] Abbasimehr H. and Paki R.," Prediction of COVID-19 Confirmed Cases Combining Deep Learning methods and Bayesian optimization," Chaos, *Solitons&Fractals*,110511,2021.
[2] Armato A.,Fanucci L., Scilingo E.P., and De Rossi D.," Low-error digital hardware implementation of artificial neuron activation functions and their derivative," *Microprocessors and Microsystems*," 35 (6) pp.557–567,2011. http://doi.org/10.1016/j.micpro.2011.05.007
[3] Bamber S.S., and Vishvakarma T., "Medical image classification for Alzheimer's using a deep learning approach," *Journal of Engineering and Applied Science*, 70(1), 2023. https:// doi.org/10.1186/s44147-023-00211-x
[4] Banerjee K. C V., Gupta R.,Vyas K.H A., and Mishra B.," Exploring Alternatives to softmax Function," 2021.https://doi.org/10.5220/0010502000002996
[5] Bekkar A., Hssina B., Douzi S., and Douzi K., "Air-pollution prediction in smart city, deep learning approach," Journal of Big Data, 8(1), 2021. https://doi.org/10.1186/ s40537-021-00548-1
[6] Bergstra J.,Remi B.Y. B., and Balazs K., "Algorithms for Hyper-Parameter Optimization," Advances in Neural Processing System 24," 2011.
[7] Chandra M., "Hardware Implementation of Hyperbolic Tangent Function using Catmull-Rom spline interpolation," arXiv (Cornell University)2020. https://doi.org/10.48550/arxiv.2007.13516
[8] Cetin O., Temurts F., and Gulgonul E., "An application of multilayer neural network on hepatitis disease diagnosis using approximation of Sigmoid function," Dicle Medical Journal, 42(2), 2015.https://doi.org/10.5798/diclemedj.0921.2015.02.0550
[9] Clevert D.A., Unterthiner T., and Hochreiter S., "Fast and accurate deep network learning by Exponential Linear Units (ELUs)," *arXiv:1511.07*289,2015. https://doi.org/10.48550/ arXiv.1511.07289
[10] De Ryck T., Lanthaler S., and Misshra S., "On the approximation of functions by tanh neural networks," *Neural Networks*, 143,732-750, 2021.https://doi.org/10.1016/ j. neunet.2021.08.015
[11] DeCastro-García N., Muñoz Castañeda N. L., Escudero García D., and Carriegos M. V.," Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a machine learning algorithm," complexity,116,2019.https://doi.org/10.1155/2019/6278908
[12] Drewil G. I., and Al-Bahadili R. J., "Air pollution prediction using LSTM deep learning and meta heuristics algorithms," *Measurement:Sensors*, 100546,2022
[13] Dubey S.R., Singh S.K., and Chaudhuri B.B.,"Activation functions in deep learning: A comprehensive survey and benchmark,"*Neuro computing*, *503*,92108,2022.https://doi.org/10.1016/ j.neucom.2022.06.111
[14] Effrosynidis D., Spiliotis E., Sylaios G., and Arampatzis A.,"Time series and regression methods for univariate environmental forecasting: An empirical evaluation," *Science oftheenvironment*,162580,2023.
[15] Global Development. Oxford Martin School, 2023.https://www.oxfordmartin.ox.uk/global-development
[16] Han Y., and Meng S.,"Machine English Translation Evaluation System Based on BP Neural Network Algorithm," *Computational Intelligence and Neuroscience*, 1–10, 2022.
[17] Hochreiter S., "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, *06*(02),107–116,1998. https://doi.org/10.1142/s0218488598000094
[18] Hochreiter S., and Schmidhuber J.," Long Short-Term Memory," Neural Computation, 9(8), 1735-1780, 1997.
[19] Hu Z., Zhang J., and Ge Y.," Handling Vanishing Gradient Problem Using Artificial Derivative," IEEE Access,9,22371-22377,2021. https://doi.org/10.1109/ access.2021.3054915
[20] Jardim R., and Morgado-Dias F., "A new solution to the hyperbolic tangent implementation in

hardware: polynomial modeling of the fractional exponential part," *Neural Computing and Applications*, 23 (2) (2013) pp.363–369.https://doi: 10.1007/s00521-012-0919-0

[21] Kunc V., and Kléma J.," Three Decades of Activations: A Comprehensive Survey of 400 Activation Functions for Neural Networks," *arXiv (Cornell University)*,2024.https://doi.org/10.48550/ arxiv.2402.09092

[22] Kaundal R., Kapoor A.S., and Raghava G.P.," Machine learning techniques in disease forecasting: a case study on rice blast prediction," BMC *Bioinformatics*, *7*(1), 485, 2006.https://doi.org/10.1186/1471-2105-7-485

[23] Li M., Jiang Y., Zhang Y., and Zhu H.," Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, 2023. https://doi.org/10.3389/ fpubh.2023.1273253

[24] Lessmann S., Stahlbock R., and Crone S.F.," Optimizing hyperparameters of support vector machines by genetic algorithms," Proc. 2005 Int. Conf. Artif. Intell. ICAI'05. 1(2005) 74–80

[25] Lupon J., Gaggin H. K., de Antonio M., Domingo M., Galan A., Zamora E., Vila J., Penafiel J., Urrutia A., Ferrer E., Vallejo N., Januzzi J. L., and Bayes-Genis A.,"Biomarker-assist score for reverse remodeling prediction in heart failure: The ST2-R2 score," International Journal of Cardiology, *184*, 337–343,2015. https://doi.org/10.1016/ j.ijcard.2015.02.019

[26] Mahaur B., Mishra K., and Singh N.,"Improved Residual Network based on norm preservation forVisual recognition," NeuralNetworks,157,305322,2023.https://doi.org/10.1016/j.neunet.2022.10.023

[27] Muhammed M. O., "Hyperparameter Optimization of a Parallelized LSTM for Time Series Prediction," Vietnam Journal of Computer Science, 1– 26, 2023.https://doi.org/10.1142/s219688882

[28] Nogueria F.,"A Python implementation of global optimization with Gaussian processes," GitHub,2024.https://github.com/fmfn/BayesianOptimization

[29] Prasanth S., Singh U., Kumar A., Tikkiwal V. A., and Chong P. H., "Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach," Chaos, Solitons & Fractals, *142*, 110336, 2021.https:// doi.org/10.1016/j.chaos.2020.110336

[30] Park S., Bong K., Shin D., Lee J., Choi S., and Yoo, H. J., "4.6 A1.93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications," *IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3.
http://doi.org/ 10.1109/ISSCC.2015.7062935

[31] Ramachandran P., Zoph B., and Le Q.V., "Searching for activation functions," *arXiv*,2017. https://doi.org/10.48550/arXiv.1710.05941

[32] Schlessman, J., "Approximation of the sigmoid function and its derivative using a minimax approach," *Theses and Dissertations*, 2002.

[33] Snoek J.,Larochelle H., and Adams R.P.,"Practical Bayesian optimization of machine learning algorithms," Advances in Neural Information Processing Systems 25, 2012.

[34] Timmons N.G., and Rice, A.,"Approximating Activation Functions," *arXiv(Cornell University)*, 2020. https://doi.org/10.48550/arxiv.2001.06370

[35] Temurtas F.,Gulbag A., and Yumusak N.," A study on Neural Network using Taylor series Expansion of sigmoid activation function," In lecture notes in computer science,pp.389-397,2004.https://doi.org/ 10. 1007/978-3-540-24768-5_41

[36] Tirupati G.,Krishnam Prasad M.H.M., & Srinivasa Rao P.,"COVID-19 Prediction Modeling Using Bidirectional Gated Recurrent Unit Network Model," Journal of Webology, Volume18(N0.5),1541,2021.https://www.webology.org/abstract.php?id=1452

[37] Tiwari V., and Khare N., "Hardware implementation of neural network with Sigmoidal activation functions using CORDIC," *Microprocessors & Microsystems* 39 (6) (2015) 373–381.

[38] Vincent P., de Brebisson A. and Bouthillier X.," Efficient exact gradient update for training deep networks with very large sparse targets," NIPS, 2015.
[39] Wei L., Cai J., Nguyen V., Chu J., and Wen K.," P-SFA: Probability based Sigmoid Function Approximation for Low complexity Hardware Implementation," Microprocessor and micro systems,76, 103105,2020.https://doi.org/10.1016/j. micpro.2020.103105
[40] Wang S.,Liu B., and Liu F.,"Escaping the Gradient Vanishing: Periodic Alternatives of Softmax inAttentionMechanism," *IEEE Access*, *9*,168749168759,2021.https://doi.org/10.1109/access.2021.3138201
[41] Willmott C., and Matsuura K.,"Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate Research*, *30*, 79–82, 2005. https:// doi.org/10.3354/cr030079
[42] Wu J., Chen X-Y., Zhang H., Xiong L-D., Lei H., & Deng S.-H.," Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization," Journal of Electronic Science and Technology, *17*(1), 26–40, 2019.https://doi.org/10.11989/ JEST. 1674-862X.80904120
[43] Zhang J., He T., Sra S., and Jadbabaie A., "Why gradient clipping accelerates Training: A theoretical justification on adaptivity," international conference learning representation,2020.
[44] Zamanlooy B., and Mirhassani M.,"An Analog CVNS-Based Sigmoid Neuron for Precise Neurochips," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *25*(3), 894–906.
[45] Zaki P.W., Hashem A.M.,Fahim E.A.,Mansor M.A.,Eigenk S.M.,Mashaly M., and Ismail S.M.,"A Novel Sigmoid Function Approximation Suitable for Neural Networks on FPGA,"15th International computing EngineeringConference,2019.https://doi.org/10.1109/icenco48310.2019.9027479
[46] G.Tirupati,MHM Krishna Prasad and P.Srinivasa Rao ,"An Improved Parallel Heterogeneous Long Short-Term Memory Model with Bayesian Optimization for Time Series Prediction", International Journal of Experimental Research and Review, ,45, 106-118,November2024.doi.org/10.52756/ijerr.2024.v45spl.009