International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Analysis of Large SARS-CoV-2 Data using Scalable Genetic Algorithm with Enhanced Bi-LSTM Method

Upendra Singh[1, *] Dr. Ajay Raundale[2]

Submitted: 23/04/2023 **Revised**: 25/06/2023 **Accepted**: 01/07/2023

Abstract: Corona Virus Disease 2019 (COVID-19), caused by the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) virus, which emerged in late 2019, is now spreading rapidly throughout the world and has reached most countries. Knowledge has led to researching the outbreak's spread and growth. The increase in cases can lead to an increase in the size of SARS-CoV-2 datasets. The scalable model needs to be developed to handle very large SARS-CoV-2 datasets. This paper proposes a scalable machine learning algorithm for huge SARS-CoV-2 prediction. A Scalable Susceptible–Infected (SSI) model consisting of a Scalable Genetic Algorithm with enhanced bi-LSTM is proposed to predict coronavirus disease using a big data framework. The experimental results of epidemic data from several cities indicate that people infected from SARS-CoV-2 show more infection in the latter part of the first week from getting infected; this has relevance to the epidemic's transmission rate. In addition, relative to conventional models for the epidemic, the hybrid model will substantially reduce prediction results errors and achieve better average absolute percentage errors (MAPEs).

Keywords: *COVID-19, SARS-CoV-2, Scalable model, Big Data, Genetic Algorithm (GA)*

1. Introduction

Global economic growth has led to the pandemic named COVID -19. From now on, there will be no treatment or vaccination for society. Life and the economy are in danger. Artificial intelligence plays a vital role in COVID -19, and the increasing cases in the pandemic combating artificial intelligence have become very important (Devaraj et al., 2021). Until effective vaccines become available, only by suppressing community outreach and following the same strict public health measures developed and implemented during SARS (Wilder-Smith et al., 2020). The number of deaths worldwide is minimized due to handwashing, social isolation, and wearing masks. Besides, the health authorities are also responsible for monitoring the current status and frequency of (Lai et al., 2020) outbreaks. At the same time, the public should take appropriate measures to cooperate.

A unified health approach can be introduced to reduce the risk of pandemic diseases and problems in the relationship between humans, animals and the environment. Many efforts are jointly made by veterinary medicine, and the intervention by humans, a multidisciplinary, holistic health approach can solve complex problems (El Zowalaty and Jarhult, 2020). But people¨ have technological weapons, and their wise use mayturn the tide of this epidemic. Devaraj et al. (2021) provides good detail of the techniques related to deep learning, such as generative adversarial networks, extreme machine learning and long-term short-term memory (LSTM), and AI-based platform challenges for COVID-19. Jamshidi et al. (2020) explains in detail the platform data provided by AI-driven.

The traditional epidemic model analyzes the infection rate, and the infected people are associated with the dynamic change. From that, the ratio of the spread of disease and the number of people infected by the pandemic. Are known. From the model, we can say that the people are infected at the same level and from the results, the trend can be known, which is a limitation. The epidemic can be controlled by using control and prevention measures; the attention can be increased by implementing control and prevention measures. Awareness can speed up the control of the virus. Epidemiological data alone is not enough to make accurate predictions. We need to use public health emergency data to model epidemics. Using the message generation function, the control and the prevention of government strategies can be done by working on the limitations of the prediction accuracy and the single factor model for the epidemic, which is traditional (Zheng et al., 2020).

Epidemiological models are used when certain diseases spread all over the glove and emergency is related to health. From this, the control and prevention measures are a subject of study. The three models largely used are the susceptible–exposed–infected–recovered (SEIR) and SI

[1]*Research Scholar, Department of Computer Science & Engineering*
[2]*Research Supervisor, Department of Computer Science & Engineering*
[1,2]*Dr. A.P.J Abdul Kalam University, Indore, MP, India, 452016*
E-mail: Upendrasingh49@gmail.com

recovered (SIR). Susceptible infection (SI), models (Kermack and McKendrick, 1927; Li et al., 1999; Yang et al., 2020a) are used as generally accepted epidemic models, and the susceptible people are shown by *S*,*E*, *I* and *R*. From this, we can get the data about the infection cases, no recovered people and the no of people incubation. The differential equation is used to know *I* and *S* was using the SEIR, SIR, SI Ebola and SARS were successfully predicted using these models as it has good capabilities for predicting disease (Berge et al., 2017; Rizkalla et al., 2007; Ng et al., 2003; Small et al., 2004; Zakary et al., 2016).

In Wuhan, China, especially in December 2019, there was an outbreak of persistent coronavirus disease (COVID-19) caused by a new virus that has not been found in humans before and is spreading rapidly and widely. Of this epidemic around the world, the number of confirmed cases is increasing rapidly every day. The number of suspicious cases is increasing according to the symptoms associated with the disease. Unfortunately, the number of deaths is also increasing. On a global scale, it is becoming more difficult to handle all the Information about these cases in different situations. When patients are injured or suspect they have symptoms, it is necessary to develop a scalable model that can be used to analyze and predict the large-scale SARS-CoV-2 disease.

Big data structures (Oussous et al., 2018) are needed to develop scalable algorithms to process big data generated from various sources. In recent years, countless processing systems have been developed specifically for big data. The advantages of MapReduce are scalability and enhanced adaptability. Veiga et al. (2016) conducted exploratory tests on Spark, Hadoop and Flink. They created resource usage for MapReduce (Li et al., 2016). Although it can be customized for non-critical errors in MapReduce, it has a multi-level in-memory programming model (Jha et al., 2020).

COVID-19 cases achieve their peak value in a short time, which resembles waves, like the first wave, second wave, and so on. So, utilization of such a huge chunk of data generated over a short time may make the situation better by predicting the next upcoming waves with low Errors in prediction and giving ideas to health framework managers, the pharmaceutical industry, the wellness industry, and even the daily necessity for the industry to handle uncertain situations motivate the introduction of big data into data generation in the COVID-19 era. The proposed SSI model is composed of GA and bi-LSTM and can be scaled through a big data infrastructure to handle large SARSCoV-2 datasets. Therefore, The prediction accuracy and the infection rate can be improved using a network module with bidirectional long short-term memory (bi-LSTM) (Shahid et al., 2020). Compared with the traditional recurrent neural network (Mikolov et al., 2010), LSTM can better record the long-term dependence of the sequence, so it is suitable for classifying and processing data with long sequences (Greff et al., 2016; Gers et al., 1999; Cho et al., 2014a,b) And forecast. In addition, a genetic algorithm (GA) is integrated, which can extract features from a large SARS-CoV-2 data set (Kazemi et al., 2014). It is used to select important features from a set of random features and then determine the applicability according to the correlation and chi-square test (Gajawada, 2019). This paper analyzes the infection rate of coronavirus in humans, which can infect more people with time and has given rise to the susceptible infection (SSI) model. The large size of the changing pattern COVID-19 data poses challenges in the proposed work's experimentation, as data drift is a common factor that affects the model and leads to an increase in false positives. These challenges were addressed by maintaining the genetic algorithm, whose selection component avoids data shift by selecting a proper subset of the database for the confidence interval of the sample used in training. The two models used for the prediction of SARS-CoV-2 are the bi-LSTM and the Scalable GA model. The accuracy of results is found in the SSI model seen from the epidemiological study results; the performance is found to be stable compared to the traditional method.

The research in this article is given as follows. The primary Information is given in the 2 section. The proposed algorithm is discussed in the 3 section. In the 4 section, the experimental results are discussed in the 5 section, and the conclusion is given.

2. Preliminaries

This section presents a methodology of the various techniques used in our proposed work.

2.1. Genetic Algorithm (GA) for feature selection

As reported by (Babatunde et al., 2014), GA is a very efficient and known method for selection because users or authors can change GA feature settings to improve its results further. The operations in genetic algorithms are an iterative process that manipulates chromosomes (decision candidates) to create new populations through genetic functions such as crossover and mutation. Jiang et al. (2017) proposed feature selection based on an improved genetic algorithm combined with a previously trained deep neural network to predict outpatient needs. The prediction model gives the deep feed-forward neural network. An initial parameter set is generated through a compound pre-learning process based on an auto-encoder to overcome the optimization problem when constructing a deep architecture. Most studies use a typical GA with a normal configuration (Kazemi et al., 2014; Liu et al., 2008; Urraca et al., 2015). The workflow of the Genetic Algorithm is shown in Fig. 1. This figure explains how a

genetic algorithm is used to select important features from all features, initialize the set with a random set of features, and then find fitness based on correlation and the chi-square test. After that, choose the top fitness value to develop new features by combining low fitness valued features. Then mutate these features, and append all features selected into the feature set if all features are explored, stop if not, then repeat.

2.1.1. Representation and initialization

0-1, the basic binary presentation is used here (Ghareb et al., 2016). The features of the subset are given by encoding (that is, each solution in the search space) and given by binary as an n-bit chromosome, and is given by the value ”1” indicates that the corresponding feature is selected, and ”0” indicates that it is not selected. The functions available are given by *n*. A good solution can be affected by the population. The initialization risk can be reduced, Ghareb et al. (2016) and the potential chromosomes were introduced. For quality improvement and convergence speed of the final solution based on the results of the filtering method.

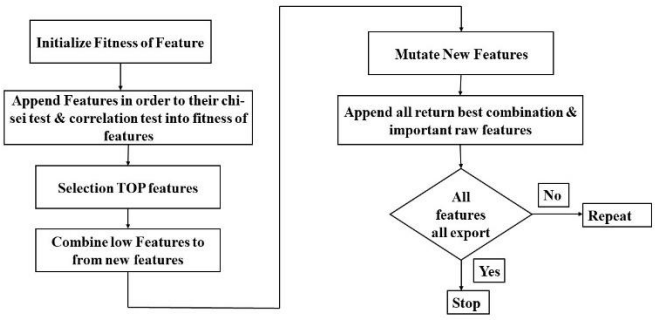

Fig 1: Workflow of Genetic Algorithm for proposed work.

2.1.2. Fitness evaluation

The main chromosome and the crossover operator are used to estimate fitness. The lowest fitness score can be used to reduce the problem. The score is more likely to be selected as a hybrid. In this study, physical fitness was evaluated by calculating physical fitness. A limited DNN model is used to accurately predict a specific subset of features and perform size loss on the subset feature. Equation one gives the fitness function 1.

$$fitness_{ci} = MS\ E_{ci} + \lambda S_{ci}. \quad (1)$$

From 1, *ci* equation, the selected chromosome is given by S_{ci} (i.e. a subset of features), Sci gives the features of the subset, and the penalty weight is given by λ. $MS\ E_{ci}$ (i.e. root mean square error) is used for the prediction of the DNN performance the cross-validation method is used for the performance. There are 3 levels and twenty nodes for every layer, and the time of computation training DNN is reduced (Jiang et al., 2017).

2.1.3. Parent selection strategy

After fitness is assessed, the applied selection strategy is used to select parents based on the relative fitness of the chromosomes. In this study, tournament selection was used to interact with noisy fitness functions and control selection pressure (Miller et al., 1995). Among *T* participants (i.e. chromosomes), *T* represents the size of the competition. The tournament winner is the chromosome with the lowest fitness score among *T* participants. The selection pressure (the degree of preference of the best individual) decreases as *T* decreases (Miller et al., 1995). In this study, two sets of tournament pools are randomly generated from the population, and each group is composed of *T* chromosomes. Considering the noise introduced by the fitness function, *T* is set to *S*/4, and *S* represents the number of chromosomes in the population. Two winners will be selected from the two pools to prepare for the next crossover operation. Compared with roulette wheel selection, this tournament draft improves the efficiency of selection by evaluating half of the participants instead of calculating the probability that each participant will be selected.

2.1.4. Crossover

Crossover operators significantly impact the quality and diversity of the next generation. GA-based FS usually uses onepoint or two-point intersection operators. Create intersection points with a fixed probability, and exchange the corresponding fragments of the parental chromosomes for creating new chromosomes. This purposeless search strategy may affect your convergence speed. MGA introduced and revised the crossover algorithm based on applicability to solve this FS problem.

2.1.5. Replacement

After the crossover, MGA uses a fixed exchange strategy. In this extremely powerful replacement, the newborn chromosome produced by the parent’s double chromosome happens to replace a person with an above-average fitness score. A simple version of the generational change, namely smooth exchange, leads to a significant increase in convergence speed, especially in the case of

multi-purpose optimization problems (Chafekar et al., 2003). To maintain population diversity in the iterative process, two constraints are added before the new chromosome replaces the old chromosome: (a) the fitness score of the new chromosome is lower than the average level; (b) the new chromosome is different from everyone in the current population Individuals. If these two restrictions can not be met, the new chromosomes are discarded, and the crossover is repeated until a matching chromosome is obtained.

2.1.6. Final feature subset

Once the MGA reaches convergence, the average fitness value changes the least within a few generations, and one or more chromosomes are selected and decoded into the final feature subset. The integration strategy is introduced into the MGA decoding operation to improve the robustness of the endpoint. First, extract the first *k* chromosomes with the lowest fitness value from the last population, where *k* is a predetermined parameter. Then, a new chromosome is created by combining all the genes on these *k* chromosomes. It is obtained by decoding this newly created chromosome. The final subset of traits produced by the combination is based on all traits selected from each of the first *k* chromosomes. Therefore, information about the combination of traits carried by the first *k* chromosomes is inherited and re-combined in the newly generated chromosomes, and the stability of the last subset of traits is increased.

2.2. Bidirectional long short-term memory (bi-LSTM)

SARS-CoV-2, responsible for infecting billions of people and economies worldwide, requires detailed research on subsequent trends to develop suitable short-term prediction models to predict the number of future cases. Hence, developing strategic plans in the public health system for death prevention and patient management needs to be carried out. Various predictive models have been proposed, including autoregressive integrated moving average (ARIMA), support vector regression (SVR), long shot term memory (LSTM), bidirectional long short-term memory (bi-LSTM) to predict the time series of confirmed cases, death and recovery in ten countries hit by SARS-CoV-2 (Shahid et al., 2020). Based on proven reliability and predictive accuracy with enhanced accuracy by Shahid et al. (2020), bi-LSTM can predict pandemics for better planning and management. Wanyan et al. (2020) developed LSTM to process time-varying patient data, applied the proposed relationship learning strategy and other static functions to the endpoint, and replaced the traditional softmax layer with a SkipGram relational learning strategy to compare the similarity between the patient and the embedded result presentation. They show that the structure of the HGM classification model can be learned reliably. In the experimental results, Wanyan et al. (2020) shows that in all prediction time windows, the HGM model based on relational learning has a higher range than the two comparator models under the receiver power curve (auROC), so the significant improvements to recall.

Devaraj et al. (2021) used ARIMA, LSTM, SLSTM and PROPHET models to develop time series forecasts of SARSCoV-2 results to estimate future forecasts of confirmed, fatal and recovered cases for the time intervals specified in the model. The proposed method can be used to predict shortterm and medium-term pollution. The analysis results show that the stacked LSTM and LSTM models are closer to other research models and have the robustness to predict SARS-CoV-2 cases. Shastri et al. (2021) conducted an in-depth investigation of the pandemic in terms of information sources and developed a preliminary study to diagnose SARS-CoV-2 using the proposed Deep-LSTM integration model. Shastri et al. (2021) conducted an experiment on confirmed SARS-CoV-2 cases and deaths in India. Various classification indicators can be used to test the effectiveness of the proposed model with an error rate. For confirmed cases of SARS-CoV-2, the accuracy rate reached 97.59%, and for death cases, the accuracy rate reached 98.88%. Furthermore, studies were carried out where official data on verified instances of COVID-19 were analyzed using Spearman's correlation and real-time travel data and data on health resources. The study covered the period from January 20 to February 19, 2020 Ying et al. (2020). Five patients with fever and respiratory symptoms were hospitalized at the Fifth People's Hospital of Anyang, China, in January 2020. We included them with one asymptomatic family member to study the effects of the cluster on the wider population. All patients gave their written permission once the local institutional review board authorized the research. The medical records of each individual were carefully examined Bai et al. (2020). Everyone who needed it had a CT scan of their chest. Nasopharyngeal swabs were used in a real-time reverse transcriptase polymerase chain reaction (RT-PCR) analysis for COVID-19 nucleic acid (Novel Coronavirus PCR Fluorescence Diagnostic Kit, BioGerm Medical Biotechnology) Bai et al. (2020). To determine the shape of the epidemic curve, they included data on population mobility before and after January 23 as well as the most recent COVID-19 epidemiological data into the Susceptible-ExposedInfectious-Removed (SEIR) model. They also used an artificial intelligence (AI) technique trained on SARS data from 2003 to forecast the pandemic Yang et al. (2020b). based on researchers' study on the LSTM model; we have developed an integration of bi-LSTM and GA to predict coronavirus diseases based on the features of bi-LSTM. These features are used to predict pandemics for better Planning and management.

2.3. Big Data Framework for SARS-CoV-2 datasets

Big data is a term used to describe a large amount of structured and unstructured data that floods the business daily. However, the amount of data is not important. What matters is the work of the organization. And data. Big data can be analyzed to provide information that leads to more effective decision-making and strategic business transfer Costa et al. (2020). Hadoop is an open-source platform that uses a simple programming model to store and process big data in a distributed environment across computer groups. Extend a server to thousands of computers, each of which provides local computing and storage (Borthakur et al., 2008).

In Wuhan, China, especially in December 2019, there was an outbreak of persistent coronavirus disease (COVID-19), which is caused by a new virus that has not been found in humans before and is spreading rapidly and widely. In this epidemic around the world, the number of confirmed cases is increasing rapidly every day, the number of suspicious cases is increasing based on the symptoms associated with the disease, and unfortunately, the number of deaths is also increasing. In many situations around the world, it becomes difficult to manage all of this information in different situations. If the patient is injured or you suspect what symptoms the patient is experiencing. Therefore, there is an urgent need to create a multi-dimensional system for large-scale storage and analysis of the generated data (Elmeiligy et al., 2020a).

Khashan et al. (2020) has developed a framework to handle complex queries against the SARS-CoV-2 data set, and the coronavirus is running in large numbers worldwide while only the Big Data application and NoSQL database are running. A small amount of data through SQL or a large amount through NoSQL. Over time, the scale of SARS-CoV-2 data collected by each country/region may become larger and larger. The size of the SQL form may not be large enough to handle this size. Therefore, you should rest assured to use the NoSQL database in this case. It is recommended to use COVID-QF (comprehensive SARS-CoV-2 data warehouse with Apache Spark and HDFS) for indexing and processing large files for use in ongoing research reports. The structure has three levels. Responsible for this layer, first check the size of large or small data sets and then check the correct data set engine that matches the user's query suggestions. At the second level, the system sends requests from users for processing. The r layer of Hadoop HDFS is used for data storage and the aggregation algorithm k of MapReduce. The final level is used with the selected SQL or NoSQL engine to complete the required work. It should be noted that a vector containing the names of SQL and NoSQL engines is first created to define the database engine that will match the user's request. Mongo database query can subdivide data and reduce query time.

Benbrahim et al. (2020) adopted, developed and tested a deep transfer learning (DTL) method based on the InvolutionV3 and ResNet50 model, which is based on the Apache Spark framework based on Convolutional Neural Network (CNN), to detect COVID on chest X-rays collected from the Kaggle repository. Elghamrawy (2020) provides an in-depth discussion on the impact of deep learning and big data analysis on disease containment. In addition, a model inspired by big data-driven deep learning (DLBD-COV) is proposed, suitable for early detection of SARS-CoV-2 cases using computed tomography or X-rays. The proposed diagnostic model is based on machine learning (H2O) for scalable processing, generating engagement networks (GAN) and convolutional neural networks (CNN) and using their classification. When the H2O infrastructure is used for scalable COVID-19 classification, the experimental results emphasize the superiority of DLBD-COV. Elmeiligy et al. (2020b) launched Apache Spark's comprehensive COVID-19 (CSS-COVID) storage system to address the daily increase in COVID-19. CSS-COVID reduces the processing time and storage of daily COVID-19 data. CSS-COVID consists of three stages: insert and index, save and query. The data is divided into subsets in the insertion step, and each subset is indexed separately. The storage layer uses many storage nodes to store data, and the request layer is responsible for processing the request process. The effectiveness of daily processing of data on injuries caused by the coronavirus. We use the Hadoop framework to make our proposed SSI model scalable to handle large SARS-CoV-2 data sets. The proposed model will be explained in the next section.

3. Proposed Work

In this paper, we discuss the proposed SSI model, which uses GA based bi-LSTM approach using a Big Data framework. We have used the Hadoop environment to implement the proposed algorithm Scalable Genetic Algorithm and enhanced bi-LSTM to form the SSI model. Fig. 2 shows the architecture of the SSI model for SARS-CoV-2 prediction by using all historical data. The fundamental guideline of the SSI approach is to utilize the proportion of the quantity of new affirmed cases at time *t* to the total number of new affirmed cases throughout various time scales before time *t* to compute the contamination rate and set up a pestilence model. Moreover, the significance of various time scales to the newly affirmed cases at time t is broken down as per the forecast consequence of the model. Assembled multi-parameter factors, which decide the effect of affirmed cases at various occasions before time t on the affirmed cases at time t, are utilized in the SSI model to measure the contamination pace of tainted cases at various periods. At that point, the improved model is utilized for dissecting

the improvement law of irresistible sicknesses. Additionally, the proposed enhanced bi-LSTM network is utilized to estimate the contamination rate deviation of the epidemic model. It is merged with the proposed SSI model to appraise the number of tainted cases. To consider the impact of government control infected cases, the media's straightforward reports, and the increment in open mindfulness concerning pandemic avoidance, this article utilizes trained proposed scalable models to extract attributes from significant information on different regions and urban areas. The extracted attributes are merged with the proposed enhanced bi-LSTM approach to address the deviation of the contamination rate assessed by the SSI model, which could foresee the number of tainted cases dependent on the transmission laws and improvement pattern. The proposed scalable models are discussed in the subsequent section.

3.1. Scalable Genetic Algorithm

Algorithm 1 is implemented on Hadoop framework (Bhosale and Gadekar, 2014). Scalable Genetic Algorithms work on the parallel processing of datasets. The workflow of the Hadoop framework is shown in Fig. 3. Algorithm 1 steps are discussed as follows: In Line 1, FT is a feature table created by operation/function. It uses the columns function to take names of all features from the dataset using a list of python operations, and the type() function is used to get the data type of every feature or column. Furthermore, Line 2 defines a data file *X*, which is parallelly processed over multiple machines. In Line 3, a feature vector is created by calling map() and reduceByKey(), which runs data files for parallel processing and performs the FT function over all files. The detailed description of map and reduceByKey is given in Algorithm 2 and Algorithm 3, respectively. Simultaneously, the line 4 fitness vector is initialized. Lines 5-10 operate on all the features in the feature vector. Additionally, if the feature selected is of type categorical in Line 6, then Line 7 performs a chi-square test to detect whether the given categorical column is relevant to decide reached over the target decision. (If the chi-square value corresponds to the target column, the feature is important). In Line 9, if the feature selected is of type numerical, then perform a correlation test in Line 10 (here, lower the feature is correlated with the target column than only important). Next, the fitness vector array selects a feature from the fitness vector with low correlation and high chi-square values from lines 11-13. Furthermore, in Line 14, if feature names are missed, then go to Line 11 for unexplored feature names. Line 16 combines the feature with high correlation and low chi-square test values and performs step 2. Finally, Line 17 returns fitness values for all captured features. A mutation is performed in Line 18 for the fitness values. In the end, Algorithm 1 returns features with the best fitness values. Here, all features are processed and ready to be utilized by the proposed bi-LSTM model.

3.2. Enhanced bi-LSTM Algorithm

Essential working of the proposed enhanced bi-LSTM uses a forward and backward approach to predict the next part of sequential data, as in the number of SARS-CoV-2 cases, death, and recoveries. We have data for a particular period. A RECURRENT NEURAL NETWORK (RNN) is created and passed the data to predict new cases, deaths, and recoveries. It takes data, initialises layers of RNN, generates output count, and then sets the weights' values to minimise error. The architecture of the proposed work is shown in Fig. 4. First, we split the dataset into train and test. After feature selection, we apply a genetic algorithm to pass data with these features into a bidirectional-LSTM network for training a model. The detailed description of the proposed bi-LSTM model is given in Algorithm 4. Line 1 of Algorithm 4 initializes sequence length to 10. In Lines 2-9, random initialization of parameters h_t and c_t takes place, where h represents the hidden layer, and c represents the combination layer. Line 5 makes a hidden layer. Line 6 makes a combination layer. After that, Line 10 assigns the parameter values, where f represents forget layer, i represents the input layer, and o represents the output layer. Additionally, W, U depicts the neural network's weights. b represents the bias for the neural network. matrix mul/element wise mul performs matrix multiplication to solve weight $*$ input and weight $*$ input + b of the neural network output. Line 10 generates all layers by computing the output of every neural network layer. The sigmoid function takes one value as input and assigns another value between 0 and 1. It is streamlined and easy to operate when building a neural network model. Activation function is changed in the proposed bi-LSTM model in Line 11. The activation function used in our model is swish (Mercioni and Holban, 2020). Swish is a smooth and comprehensive operation that consistently adapts or replaces the train in deep networks that apply across various challenging domains such as classification. In intense networks, Swish achieves higher test accuracy than ReLU. Finally, the activation function is updated on Line 13. Line 14 of Algorithm 4 initializes the RNN. ht represents the hidden layer, and ct represents the combination layer of the RNN network, where RNN stands for the recurrent neural network as the model considers RNN as baseline architecture. Hence, it inherits all basic properties of the RNN networks. From Line 17-19, the first, second, and third layer of LSTM is added with the setting of dropout regularization in all three layers. Finally, the output layer of LSTM is added with dropout regularization in Line 20. Dropout is a regularization method that bypasses input to LSTM units when activating the network, activating repeated connections and performing weight updates. This has the

effect of reducing over-fitting and improving model performance. RNN is compiled in Line 21. At last, fit RNN to the training set is performed. The fit function is used to train the bilstm model, which is initialized as sequential() and imported from Keras. It is directly imported from Keras and just using it, not any changes. Sequential() is a model available on Keras in the sequential library and only imported directly from Keras and store into the regressor variable in Line 23. The model appears to be highly complex as it follows several patterns of data movement through various models and steps, making it appear to be a more time-intensive process to extract useful insights from the dataset. Still, the model's complexity is primarily determined by the performance of the genetic algorithm and RNN complexity, which appear to perform in parallel in flow, where genetic algorithms give the order of (gnm), where g is generations, n is the population size, and m is individual size. RNN complexity is O(n), so while executing n inputs in a workflow manner, both algorithms give us the order of (gn2m) as the whole model complexity.

4. Experimental Results

The experimental result compares the proposed algorithms' performance with state-of-art methods.

4.1. Dataset Description

The data set used in this paper is taken from multiple sources such as Kaggle, data world, and WHO. This data is highly stochastic as an increase/decrease in several cases depends on other environmental/physical variables.

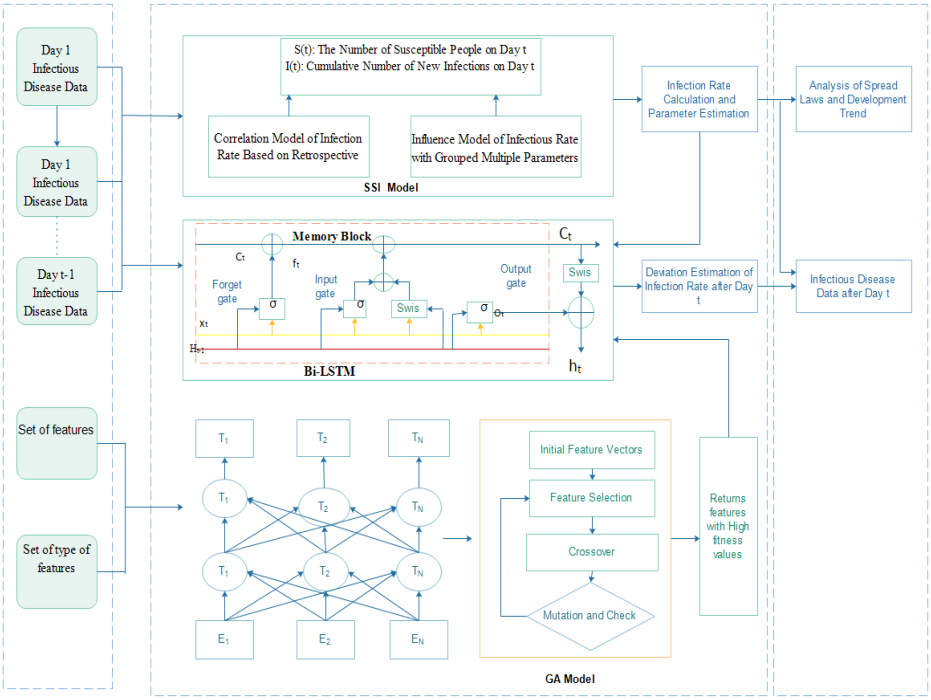

Fig 2: Workflow of Proposed SSI model

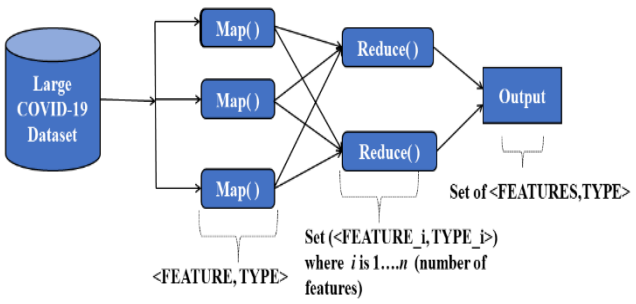

Fig 3: Integration of Hadoop framework on Genetic Algorithm

Algorithm 1: Scalable Genetic Algorithm

Data: data frame df

Result: Feature name and type with high fitness values

1 FT=list(df.columns,type(df.columns))

2 X=df

3 Feature Vector=X.map(FT).reduceByKey()

4 Fitness of features= []

5 **foreach** *features F in Feature Vector* **do**

6 **if** F = categorical **then**

7 **Apply** chi-sq test

8 **else**

9 F=numerical

10 Calculate Correlation (Michalak and Kwasnicka, 2006).

11 **foreach** *Value of Fitness of feature* **do**

12 **Select** features with low correlation value.

13 **Passed** features for chi-sq test (Gajawada, 2019).

14 **foreach** *feature not related* **do**

15 **Goto** Step 11.

16 **Combine** features and perform step 2.

17 **Return** Fitness of feature.

18 Mutate the fitness of features with mutation rate.

19 **Return** Fitness of features.

Algorithm 2: map()

Data: f, t

Result: < f, t >

1 **foreach** *feature in FT* **do**

2 **Initialize:**

3 f =feature name

4 t = data type of feature(f)

5 **Return** < f, t >

Algorithm 3: ReduceByKey()

Data: set

Result: < f, < f, t >>

1 **foreach** *feature and type tuple* **do**

2 set = set of features

3 **Return** < f, < f, t >>

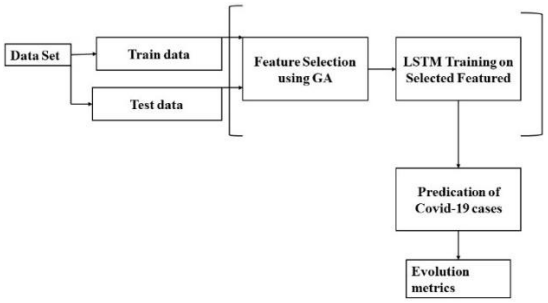

Fig 4: Architecture of proposed work

Algorithm 4: Enhanced bi-LSTM Algorithm

Data: preprocessed dataset

Result: prediction number of COVID-19 cases

Initialize sequence length = 10

foreach *i in range 0 to sequence length* **do**

    Randomly initialize h_t, c_t

    if *i=0* **then**

        ht_1 = random () ;

        ct_1 = random () ;

    else

        $ht_1 = h_t$;

        $ct_1 = c_t$;

Assign parameter values:

$f_t = sigmoid(matrix\ mul(W f, xt) + matrix\ mul(U f, ht_1) + b f)$

$i_t = sigmoid(matrix\ mul(Wi, xt) + matrix\ mul(Ui, ht_1) + bi)$

$o_t = sigmoid(matrix\ mul(Wo, xt) + matrix\ mul(Uo, ht_1) + bo)$

$cp_t = Swish(matrix\ mul(Wc, xt) + matrix\ mul(Uc, ht_1) + bc)$

c_t = element wise mul(f_t, ct_1) + element wise mul(i_t, cp_t)

h_t = element wise mul(o_t, Swish(c_t))

Initializing the RNN

regressor = Sequential()

Add LSTM layers and Dropout regularization

for *layer* = 1 **to** *layer* = 4 **do**

        regressor.add(Bidirectional(LSTM(units, sequences)))

        regressor.add(Dropout())

Add output layer: regressor.add(Dense(units))

21 Compile RNN: regressor.compile(*optimizer = adam,loss = mean squared error*)

22 Fit RNN to the training set

23 $regressor.fit(X_{train}, y_{train}, epochs = 75, batch\ size = 32)$

It consists of 32 timeseries data of confirmed SARS-CoV-2 cases in each state (28) and union territories (4) since March 14, 2020. Each series's missing values are input with the missing data statistics technique known as a linear weighted moving average. The model maintains the sequential learning ability and is feasible to produce accurate future predictions. We have taken data for study purposes for a whole year. We have split the data into 80% of it for training and 20% for testing. The Link for the dataset used is as follows:

1. CORD-19-research-challenge
2. COVID-19-in-india
3. COVID-19-state-wise-data
4. Most-common-words-in-the-cord-19-dataset
5. Coronavirus disease (COVID-19) pandemic

4.2. Performance Evaluation

4.2.1. MAPE

The presentation of our proposed expectation techniques is looked at as far as certain exhibition measure lists like mean absolute percentage error (MAPE) (De Myttenaere et al., 2016) and is depicted as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\frac{|pi - ai|}{|ai|}\,|ai| \times 100\% \qquad (2)$$

4.2.2. MAE

Mean Absolute Error (MAE) measures the average magnitude of absolute differences between *N* predicted vectors (Chai and Draxler, 2014). The MAE is depicted as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}||xi - yi||1\,| \qquad (3)$$

As discussed in earlier sections, COVID-19 was quickly generating frequent shifts in data, which represents a time series-like pattern in the data. Generally, data changes over duration are evaluated using the mean absolute error or mean squared error, as Mean fundamental percentage error is a relative error measure that uses absolute values to keep the positive and negative errors from cancelling one another out and uses relative errors to enable you to compare forecast accuracy between time-series models, These findings lead to considering MAE and MAPE to evaluate proposed work.

4.3. Results and Discussion

4.3.1. Illustrative Example

Fig. 5 explains the SARS-CoV-2 patient in India. In this figure, more red colour means more COVID patient, and less red mean fewer patients. The red colour comes under the severe zone, where states with positive SARS-CoV-2 patients above 2000 increased daily by more than 5%. The light red colour on the map indicates a moderate zone, where states come between 200 and 2000 and the daily increment is less than 5%. The light green region comes under the mild zone in All states, where the total number of positive SARS-CoV-2 cases is below 200, and a daily rise is below 2%. Low mild zone depicted by dark green colour, all the states where positive SARS-CoV-2 are below 100 and daily rise is less than 1%. Spread analysis has divided India into four categories based on the positive number of SARSCoV-2 cases and daily rise, as shown in the India map in Fig.5.

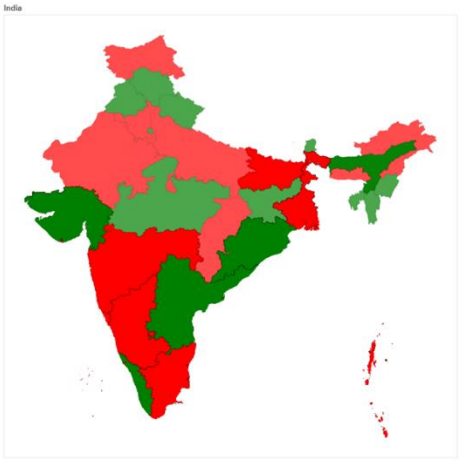

Fig 5: Division of India in the severe, moderate and mild zones depending upon the number of confirmed SARS-CoV-2 positive cases and daily rise based on the data till May 14, 2020.

4.4. Parameters used in experimental evaluation

Table 1 displays the values used by the proposed model; for specific parameters, the number of layers was 128 for hidden as in the algorithm, which represents the hidden layers; as we grow hidden from this point, the model leads to overfitting, and below this model, there is a lack of data mapping patterns, making this value best suited for that parameter. Then neurons represent the number of neurons per layer; to make the learning process effective, the count of neurons was selected as 64, which makes the bias of each node after activation nonfluctuating and makes the results stand, as well as the trade-off between variance and bias. Epochs are defined based on computing time, and a whole model architecture of 100 gives the best results. Still, increasing it leads to better performance, but only by 0.1% to 0.5%, making it the best value in terms of time and machine cost. Keeping the learning rate at 0.03 to optimize with efficiency and time effort is also considered error stabilizing at this value. The dropout layer manages overfitting with 0.47. As we increase that, it starts losing its objective of handling overfitting; con2d shows the architecture pattern of the proposed model.

Table 1: Parameters used in experimental evaluation

parameters	Value
Number of layers	128-hidden
neurons	64-per layer
epoch	100
Learning rate	0.03
Dropout rate	0.47
Con2d	4X4,4X8

4.5. Performance Prediction

The number of SARS-CoV-2 positive cases predicted statewide for months in India by the proposed scalable GAbased Feature Selection with enhanced Bi-directional LSTM model is presented in Table 2. This model is also tested statewise for daily and weekly predictions for data. Using the proposed scalable GA-based Feature Selection with an enhanced Bi-directional LSTM model, we expect several new cases, deaths, and monthly recoveries until mid-year 2021.

Table 2: Mean Absolute Percentage Error (MAPE) of states and union territories (UTs) of India by convolutional (Arora et al., 2020), stacked (Arora et al., 2020), bi-LSTM (Arora et al., 2020), and proposed scalable GA-based Feature Selection with enhanced Bi-directional LSTM model.

S. No.	States/UTs	Convolutional LSTM	Stacked LSTM	bi-LSTM	Scalable GA based Enhanced bi-LSTM
1	Andaman and Nicobar	0	0.2	0	0
2	Andhra Pradesh	3.2	1.6	1.24	1.18
3	Arunachal Pradesh	0	0	0	0
4	Assam	7.28	6.3	5.49	5
5	Bihar	7.03	4.95	5.3	5.1
6	Chandigarh	8.76	8.3	6.64	6.0
7	Chhattisgarh	12.94	11.05	10.9	10
8	Delhi	2.86	3.4	2.13	2
9	Goa	0	0	0	0

10	Gujarat	2.78	2.02	0.99	0.98
11	Haryana	5.94	5.23	4.35	4
12	Himachal Pradesh	5.57	3.81	2.68	2.28
13	Jammu and Kashmir	2.36	1.82	1.53	1.53
14	Jharkhand	5.46	3.53	2.95	2.5
15	Karnataka	3.06	2.31	1.71	1.6
16	Kerala	2.04	0.74	0.63	0.6
17	Ladakh	12.23	11.19	7.63	7
18	Madhya Pradesh	4.38	4.44	1.9	1.9
19	Maharashtra	2.43	2.23	1.29	1.20
20	Manipur	0	0	0	0
21	Meghalaya	1.1	0.55	0.55	.4
22	Mizoram	0	0	0	0
23	Odisha	7.79	6.4	5.88	5
24	Puducherry	3.13	12.65	3.13	3
25	Punjab	18.02	12.07	7.95	7.5
26	Rajasthan	1.3	2.35	1.35	1.0
27	Tamil Nadu	7.17	5.33	3.53	1.90
28	Telangana	1.83	1.39	0.97	0.8
29	Tripura	21.16	30.67	15.35	12
30	Uttar Pradesh	3.37	2.32	1.11	0.7
31	Uttarakhand	2.03	2.26	1.8	1.4
32	West Bengal	6.25	4.95	4.16	3

Table 3: Daily and weekly error percentages for one-week testing data using proposed scalable GA-based Feature Selection with enhanced Bi-directional LSTM model.

Date	Maharashtra		Madhya Pradesh		Delhi	
	Daily Error %	Weekly Error %	Daily Error %	Weekly Error %	Daily Error %	Weekly Error %
07-Mar	2.2	0.7	7.1	5.2	1.4	3.5
08-Mar	0.6	0	0.2	7.1	0	1.4
09-Mar	0.5	0.88	1.3	0.7	0.9	1.7
10-Mar	6.4	7.8	5.2	6.7	3.8	4.2
11-Mar	0.8	0.9	2.0	5.2	4.9	3.5

12-Mar	0.3	1.6	1.7	7.3	0.8	2.6
13-Mar	0.4	1.5	6.2	4.9	0.5	4.2
MAPE	0.6	1.4	0.3	1.4	1.7	0.3

In Table 3, daily and weekly prediction errors are calculated for three states of India (Maharashtra, Madhya Pradesh, and Delhi). The highly accurate, state-wise predictions will help the state-authorities balance the load that medical infrastructure can take. Depending upon predictions, several decisions can be taken, like imposition or removal of lockdowns. This would also ensure that economic activities can resume, which otherwise may create livelihood challenges for millions of people. Table 2 shows the Mean Absolute Percentage Error (MAPE) of states and union territories (UTs) of India by convolutional (Arora et al., 2020), stacked (Arora et al., 2020), bi-LSTM (Arora et al., 2020), and proposed scalable GA based feature selection with an enhanced Bi-directional LSTM model. MAPE has to calculate the 32 states in India; through the Convolutional LSTM method, MAPE is a minimum of 0 and a maximum of 21.16. The stacked LSTM method represents MAPE with a minimum of 0 and a maximum of 30.67. Through the bi-LSTM method, MAPE has a minimum of 0 and a maximum of 15.35. The Scalable GA-based Enhanced bi-LSTM method represents MAPE as having a minimum of 0 and a maximum of 12. Table 3 shows daily and weekly error percentages for one-week testing data using the proposed scalable GA-based feature selection with an enhanced Bi-directional LSTM model in Maharashtra, Madhya Pradesh, and Delhi. Its table displays the error rate between March 7 and March 13. In Maharashtra, the daily error rate is a minimum of 12 and a maximum of 10. The weekly error rate is a minimum of 8 March and a maximum of 10 March. In Madhya Pradesh, the daily error rate is a minimum of the March 8 and a maximum of the March 10. The weekly error rate is lowest on March 9 and highest on March 12. In Delhi, the daily error rate is a minimum of 8 and a maximum of 11. The weekly error rate is a minimum of the March 9 and a maximum of the March 13.

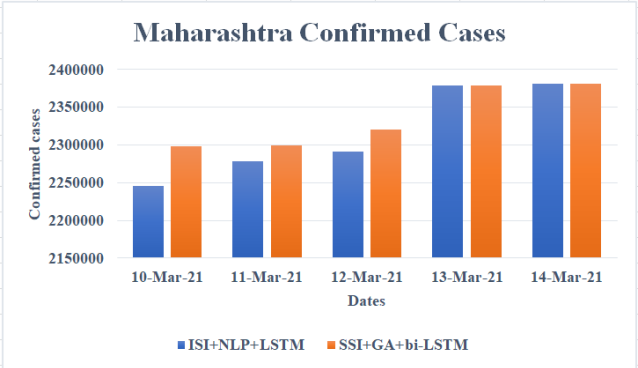

Fig 6: Maharashtra confirmed cases

4.5.1 Maharashtra Confirmed Cases

Table 4 contains information on the results obtained from several Maharashtra-confirmed SARS-CoV-2 cases from 10,11,12,13,14 March 2021. All mentioned data predicted by our proposed model SSI+SGA+Enhanced bi-LSTM (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) is predicted approx accurate confirmed cases in Maharashtra, Existing SSI+NLP+LSTM (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model predicts less confirmed cases versus actual cases to all five days. Hence, our model is better as per compare existing model. The proposed model and existing model date-wise differences for 10 march 53109 cases, 11 march 21011 cases, 12 march 29295, 13 march 27 cases, and 14 march 89 cases are different based on the existing model and proposed model. Hence, the proposed model analyses better prediction per the existing model (Zheng et al., 2020). Figure 6 represents the graphical analysis of the proposed SSI model compared with the ISI model (Zheng et al., 2020) for Maharashtra.

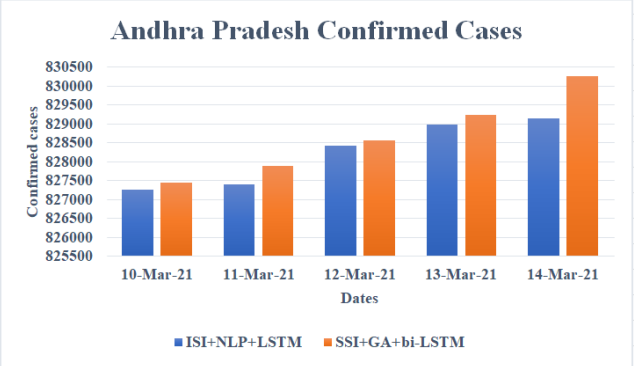

Fig 7: Andhra Pradesh confirmed cases

4.5.2. Andhra Pradesh Confirmed Cases

Table 5 contains information on the results obtained from several Andhra Pradesh confirmed SARS-CoV-2 cases from 10,11,12,13,14 March 2021. All mentioned data predicted by our proposed model SSI+SGA+Enhanced bi-LSTM (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) is expected approx accurate confirmed cases in Andhra Pradesh, Existing ISI+NLP+LSTM (Zheng et al., 2020)(Zheng et al., 2020) (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model predicts less confirmed cases versus actual patients to all 5 days. Hence our model is better as per compare existing model. The proposed model and existing model date-wise difference cases like 10 march 197, 11 march 503, 12 march 131, 13 march 254, and 14 march 1108 cases differ based on the existing model and proposed model. Hence, the proposed model is analysis better predication as per the current model. Figure 7 represents the graphical analysis of proposed SSI model compared with the ISI model (Zheng et al., 2020) for Andhra Pradesh

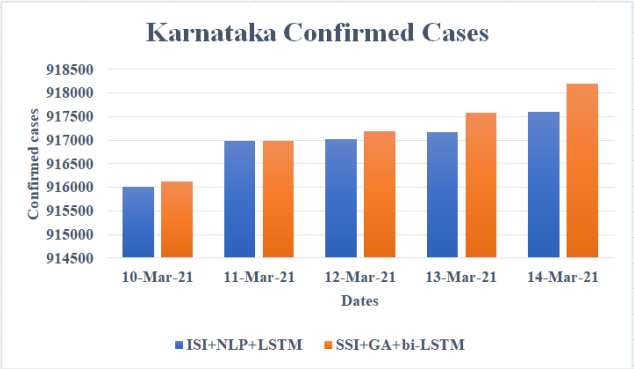

Fig 8: Karnataka confirmed cases

4.5.3. Karnataka Confirmed Cases

Table 6 contains information on the results obtained from several Karnataka-confirmed SARS-CoV-2 cases from 10,11,12,13,14 March 2021. All mentioned data predicted by our proposed model SSI+SGA+Enhanced bi-LSTM (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) is predicted approx accurate confirmed cases in Karnataka, Existing ISI+NLP+LSTM (Zheng et al., 2020) (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model predicts less confirmed cases versus actual cases to all 5 days. Hence, our model is better as per compare existing model. The proposed model and existing model date-wise difference cases like 10 march 112, 11 march 12, 12 march 163, 13 march 415, and 14 march 592 cases are different based on the existing model and proposed model. Hence, the proposed model is analysis better prediction as per the existing model. Figure 8 represents the graphical analysis of proposed SSI model compared with the ISI model (Zheng et al., 2020) for Karnataka.

Table 4: Maharashtra confirmed cases

Dates	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
10-Mar-21	2245876	2298985	53109
11-Mar-21	2278979	2299990	21011

12-Mar-21	2291070	2320365	29295
13-Mar-21	2378674	2378701	27
14-Mar-21	2380802	2380891	89
MAE	145.2	125.4	

Table 5: Andhra Pradesh confirmed cases

Dates	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
10-Mar-21	827254	827451	197
11-Mar-21	827391	827894	503
12-Mar-21	828431	828562	131
13-Mar-21	828987	829241	254
14-Mar-21	829153	830261	1108
MAE	7.5	4.1	

4.5.4. Kerala Confirmed Cases

Table 10 contains information on the results obtained from several Kerala-confirmed SARS-CoV-2 cases from 10,11,12,13,14 March 2021. All mentioned data predicted by our proposed model SSI+SGA+Enhanced bi-LSTM (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) is predicted approx accurate confirmed cases in Kerala, Existing ISI+NLP+LSTM (Zheng et al., 2020) (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model predicts less confirmed cases versus actual cases to all 5 days. Hence, our model is better as per compare existing model. In the proposed model and existing model date wise difference cases like 10 march 44, 11 march 98, 12 march 102, 13 march 53, and 14 march 189 cases are different based on the existing model and proposed model, so the proposed model is analysed better predication as per existing model. Figure 9 represents the graphical analysis of proposed SSI model compared with the ISI model (Zheng et al., 2020) for Maharashtra.

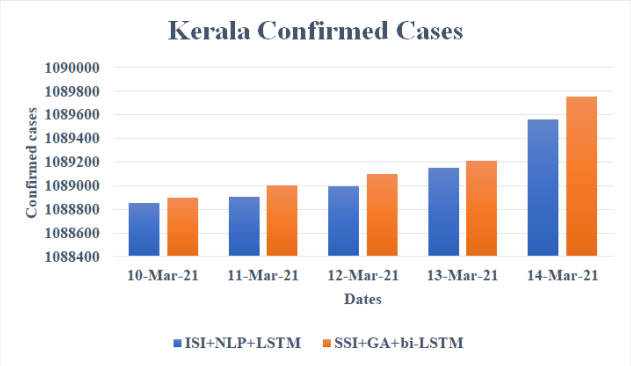

Fig 9: Kerala confirmed cases

4.5.5. Madhya Pradesh Confirmed Cases

Table 7 contains information on the results obtained from several Madhya Pradesh-confirmed SARS-CoV-2 cases from 10,11,12,13,14 March 2021. All mentioned data predicted by our proposed model SSI+SGA+Enhanced bi-LSTM (Scalable (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) is expected approx accurate confirmed cases in Madhya Pradesh, Existing ISI+NLP+LSTM (Zheng et al., 2020) (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model predicts less confirmed cases versus actual patients to all 5 days. Hence, our model is better as per compare existing model. In the proposed model and existing model date wise difference cases like

10 march 153, 11 march 101, 12 march 15, 13 march 86, and 14 march 165 cases are different based on the existing model and proposed model, so the proposed model is analysed better predication as per current model. Figure 10 represents the graphical analysis of the proposed SSI model compared with the ISI model (Zheng et al., 2020) for Maharashtra.

Table 6: Karnataka confirmed cases

Dates	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
10-Mar-21	916023	916135	112
11-Mar-21	916985	916997	12
12-Mar-21	917023	917186	163
13-Mar-21	917175	917590	415
14-Mar-21	917602	918194	592
MAE	3.75	3.29	

Table 7: Madhya Pradesh confirmed cases

Dates	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
10-Mar-21	272542	272695	153
11-Mar-21	272685	272786	101
12-Mar-21	272987	273002	15
13-Mar-21	273020	273106	86
14-Mar-21	273403	273568	165
MAE	0.69	0.58	

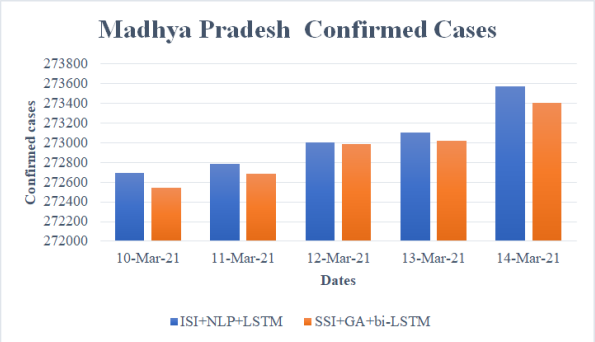

Fig 10: Madhya Pradesh confirmed cases

4.5.6. Delhi Confirmed Cases

Table 8 contains information on the results obtained from several Delhi-confirmed SARS-CoV-2 cases from 10,11,12,13,14 March 2021. All mentioned data predicted by our proposed model SSI+SGA+Enhanced bi-LSTM (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) is predicted approx accurate confirmed cases in Delhi, Existing ISI+NLP+LSTM (Zheng et al., 2020) (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model predicts less confirmed cases versus actual cases to all 5 days. Hence, our model is better as per compare existing model. The proposed model and existing model date-wise difference cases like 10 march 1192, 11 march 1144, 12 march 1737, 13 march 203, and 14 march 779 cases are different based on the existing model and proposed model. Hence, the proposed model is an analysis of better prediction per the existing model. Figure 11 represents the graphical analysis of the proposed SSI model compared with the ISI model (Zheng et al., 2020) for Maharashtra.

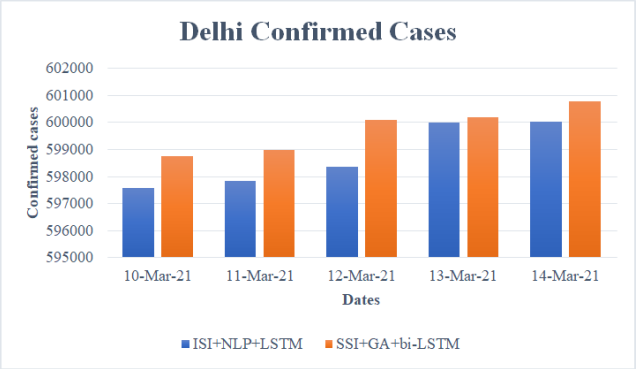

Fig 11: Delhi confirmed cases

4.5.7. State-wise Mean Squared Error (MSE)

Mean square error concerning states wise existing ISI+NLP+LSTM (Zheng et al., 2020) (improved susceptible–infected+ natural language processing+ Long Short-Term Memory) model, like Maharashtra 145.2, Andhra Pradesh 7.5, Karnataka 3.75, Kerala 308.5, Madhya Pradesh 0.69, and Delhi 4.12. Mean square error concerning states-wise proposed SSI+SGA+Enhanced bi-LSTM (Scalable Susceptible–Infected +Scalable Genetic Algorithm+ Enhanced Bidirectional Long Short-Term Memory) model, Maharashtra 125.4, Andhra Pradesh 4.1, Karnataka 3.29, Kerala 297.9, Madhya Pradesh 0.58, and Delhi 3.95. In the proposed method, all states are resulting respect to MSE is less as per compared existing method like Maharashtra 19.8, Andhra Pradesh 3.4, Karnataka 0.46, Kerala 10.6, Madhya Pradesh 0.11, and Delhi 0.17

Table 8: Delhi confirmed cases

Dates	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
10-Mar-21	597562	598754	1192
11-Mar-21	597834	598978	1144
12-Mar-21	598365	600102	1737
13-Mar-21	600000	600203	203
14-Mar-21	600010	600789	779
MAE	4.12	3.95	

Table 9: State-wise MSE

States	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
Maharashtra	145.2	125.4	19.8
Andhra Pradesh	7.5	4.1	3.4
karnataka	3.75	3.29	0.46
Kerala	308.5	297.9	10.6
Madhya Pradesh	0.69	0.58	0.11
Delhi	4.12	3.95	0.17

Table 10: Kerala confirmed cases

Dates	ISI+NLP+LSTM	SSI+SGA+Enhanced bi-LSTM	Difference
10-Mar-21	1088856	1088900	44

11-Mar-21	1088904	1089002	98
12-Mar-21	1089000	1089102	102
13-Mar-21	1089156	1089209	53
14-Mar-21	1089563	1089752	189
MAE	308.5	297.9	

5. Conclusion

This article aims to assess the trend of COVID 19 newly confirmed cases were found on different days. Intervals may have different contributions to upcoming infections. Impact of cases confirmed in recent days T analyzes the timing of new daily confirmed cases. Based on this, we propose a group multiparameter strategy it determines the infection rate of previously diagnosed cases in different groups depending on the time. And then, we suggested an SSI model with multiple parameters. This article uses scalable GA technology to analyze the collection of relevant news information using Big Data can be done, Infection control measures and habitat awareness for infectious disease prevention are meaningfully encoded features. Then, these features are fed into the enhanced bi-LSTM network to Update the infection rate provided by the SSI model. In short, it is based on the SSI model, a scalable machinelearning model built to assess the SARS-CoV-2 indicated in this article. The scalable GA is introduced to make the module efficient, which can produce results in less amount of time. The actual infectious disease cases are proving to be specific in the SSI model, which accurately analyzes the law of transmission of Viral development trends compared to previously developed models. Other than that, We provide an effective method for transmission evaluation law and development trends in public health events in the future. The proposed work poses the following future scope of improvements:

- Data drift as COVID-19 changes its spread pattern makes the model confusing when explored over any uncertain pandemic as the variance of the spread changes. These data shifts create the future scope to perform statistical tests to analyse the dataset before passing it into models, which normalise the problem statement as pandemic analysis using a high generation rate of the dataset.
- Because optimization is performed at the modular level rather than the structural level, there is room for improvement in the proposed model's internal mechanisms and working workflow.

6. Declarations

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Author Contributions

Mr. Upendra Singh performed the research work under the supervision of Dr. Ajay R. Raundalea. This manuscript is written by Mr. Upendra Singh under the guidance of Dr. Ajay R. Raundalea. This manuscript is reviewed and proofread by Dr. Ajay R. Raundalea.

Funding

No funding was received from any organization for conducting the study of the submitted work and preparation of this manuscript.

Conflict of interest

Author A and B declare that he has no conflict of interest.

Informed Consent

The research papers which are used for the study of the submitted work have been cited in the manuscript and the details of the same have been included in the reference section.

Acknowledgments

The authors would like to thank the monkeypox patient who took part in this study and also thanks the Kaggle organization for the dataset provided for our research.

Competing interests

Not always applicable and includes interests of not a financial or personal nature

Availability of data and materials

Availability as per request

References

[1] Arora, P., Kumar, H., Panigrahi, B.K., 2020. Prediction and analysis of covid19 positive cases using deep learning models: A descriptive case study

of india. Chaos, Solitons & Fractals 139, 110017.

[2] Babatunde, O.H., Armstrong, L., Leng, J., Diepeveen, D., 2014. A genetic algorithm-based feature selection .

[3] Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.Y., Chen, L., Wang, M., 2020. Presumed asymptomatic carrier transmission of covid-19. Jama 323, 1406– 1407.

[4] Benbrahim, H., Hachimi, H., Amine, A., 2020. Deep transfer learning with apache spark to detect covid-19 in chest x-ray images. Romanian Journal of Information Science and Technology 23, S117–S129.

[5] Berge, T., Lubuma, J.S., Moremedi, G., Morris, N., Kondera-Shava, R., 2017. A simple mathematical model for ebola in africa. Journal of biological dynamics 11, 42–74.

[6] Bhosale, H.S., Gadekar, D.P., 2014. A review paper on big data and hadoop. International Journal of Scientific and Research Publications 4, 1–7. Borthakur, D., et al., 2008. Hdfs architecture guide. Hadoop Apache Project 53, 2.

[7] Chafekar, D., Xuan, J., Rasheed, K., 2003. Constrained multi-objective optimization using steady state genetic algorithms, in: Genetic and Evolutionary Computation Conference, Springer. pp. 813–824.

[8] Chai, T., Draxler, R.R., 2014. Root mean square error (rmse) or mean absolute error (mae). Geoscientific Model Development Discussions 7, 1525–1534.

[9] Cho, K., Van Merrienboer, B., Bahdanau, D., Bengio, Y., 2014a. On the prop-¨ erties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 .

[10] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F.,¨ Schwenk, H., Bengio, Y., 2014b. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 .

[11] Costa, J.P., Grobelnik, M., Fuart, F., Stopar, L., Epelde, G., Fischaber, S., Poliwoda, P., Rankin, D., Wallace, J., Black, M., et al., 2020. Meaningful big data integration for a global covid-19 strategy. IEEE Computational Intelligence Magazine 15, 51–61.

[12] De Myttenaere, A., Golden, B., Le Grand, B., Rossi, F., 2016. Mean absolute percentage error for regression models. Neurocomputing 192, 38–48.

[13] Devaraj, J., Elavarasan, R.M., Pugazhendhi, R., Shafiullah, G., Ganesan, S., Jeysree, A.K., Khan, I.A., Hossain, E., 2021. Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant? Results in physics 21, 103817.

[14] El Zowalaty, M.E., Jarhult, J.D., 2020. From sars to covid-19: A previously un-¨ known sars-related coronavirus (sars-cov-2) of pandemic potential infecting humans–call for a one health approach. One Health 9, 100124.

[15] Elghamrawy, S., 2020. An h 2 o's deep learning-inspired model based on big data analytics for coronavirus disease (covid-19) diagnosis, in: Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach. Springer, pp. 263–279.

[16] Elmeiligy, M.A., Desouky, A.I.E., Elghamrawy, S.M., 2020a. A multidimensional big data storing system for generated covid-19 large-scale data using apache spark. arXiv preprint arXiv:2005.05036 .

[17] Elmeiligy, M.A., Desouky, A.I.E., Elghamrawy, S.M., 2020b. A multidimensional big data storing system for generated covid-19 large-scale data using apache spark. arXiv preprint arXiv:2005.05036 .

[18] Gajawada, S., 2019. Chi-square test for feature selection in machine learning.

[19] Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: Continual prediction with lstm .

[20] Ghareb, A.S., Bakar, A.A., Hamdan, A.R., 2016. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications 49, 31–47.

[21] Greff, K., Srivastava, R.K., Koutn´ık, J., Steunebrink, B.R., Schmidhuber, J., 2016. Lstm: A search space odyssey. IEEE transactions on neural networks and learning systems 28, 2222–2232.

[22] Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., Jamshidi, M., La Spada, L., Mirmozafari, M., Dehghani, M., et al., 2020. Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. IEEE Access 8, 109581–109595.

[23] Jha, P., Tiwari, A., Bharill, N., Ratnaparkhe, M., Mounika, M., Nagendra, N., 2020. A novel scalable kernelized fuzzy clustering algorithms based on inmemory computation for handling big data. IEEE Transactions on Emerging Topics in Computational Intelligence .

[24] Jiang, S., Chin, K.S., Wang, L., Qu, G., Tsui, K.L., 2017. Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient

department. Expert systems with applications 82, 216–230.

[25] Kazemi, S., Seied Hoseini, M.M., Abbasian-Naghneh, S., Rahmati, S.H.A., 2014. An evolutionary-based adaptive neuro-fuzzy inference system for intelligent short-term load forecasting. International transactions in operational research 21, 311–326.

[26] Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character 115, 700–721.

[27] Khashan, E.A., Eldesouky, A.I., Fadel, M., Elghamrawy, S.M., 2020. A big data based framework for executing complex query over covid-19 datasets (covid-qf). arXiv preprint arXiv:2005.12271 .

[28] Lai, C.C., Shih, T.P., Ko, W.C., Tang, H.J., Hsueh, P.R., 2020. Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and coronavirus disease2019 (covid-19): The epidemic and the challenges. International journal of antimicrobial agents 55, 105924.

[29] Li, M.Y., Graef, J.R., Wang, L., Karsai, J., 1999. Global dynamics of a seir model with varying total population size. Mathematical biosciences 160, 191–213.

[30] Li, R., Hu, H., Li, H., Wu, Y., Yang, J., 2016. Mapreduce parallel programming model: a state-of-the-art survey. International Journal of Parallel Programming 44, 832–866.

[31] Liu, Y., Yin, Y., Gao, J., Tan, C., 2008. Wrapper feature selection optimized svm model for demand forecasting, in: 2008 The 9th International Conference for Young Computer Scientists, IEEE. pp. 953–958.

[32] Mercioni, M.A., Holban, S., 2020. P-swish: Activation function with learnable parameters based on swish activation function in deep learning, in: 2020 International Symposium on Electronics and Telecommunications (ISETC), IEEE. pp. 1–4.

[33] Michalak, K., Kwasnicka, H., 2006. Correlation-based feature selection strategy in neural classification, in: Sixth international conference on intelligent systems design and applications, IEEE. pp. 741–746.

[34] Mikolov, T., Karafiat, M., Burget, L.,´ Cernock˘ y, J., Khudanpur, S., 2010. Re-` current neural network based language model, in: Eleventh annual conference of the international speech communication association.

[35] Miller, B.L., Goldberg, D.E., et al., 1995. Genetic algorithms, tournament selection, and the effects of noise. Complex systems 9, 193–212.

[36] Ng, T.W., Turinici, G., Danchin, A., 2003. A double epidemic model for the sars propagation. BMC Infectious Diseases 3, 1–16.

[37] Oussous, A., Benjelloun, F.Z., Lahcen, A.A., Belfkih, S., 2018. Big data technologies: A survey. Journal of King Saud University-Computer and Information Sciences 30, 431–448.

[38] Rizkalla, C., Blanco-Silva, F., Gruver, S., 2007. Modeling the impact of ebola and bushmeat hunting on western lowland gorillas. EcoHealth 4, 151–155.

[39] Shahid, F., Zameer, A., Muneeb, M., 2020. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. Chaos, Solitons & Fractals 140, 110212.

[40] Shastri, S., Singh, K., Kumar, S., Kour, P., Mansotra, V., 2021. Deep-lstm ensemble framework to forecast covid-19: an insight to the global pandemic. International Journal of Information Technology , 1–11.

[41] Small, M., Shi, P., Tse, C.K., 2004. Plausible models for propagation of the sars virus. IEICE transactions on fundamentals of electronics, communications and computer sciences 87, 2379–2386.

[42] Urraca, R., Sanz-Garc´ıa, A., Fernandez-Ceniceros, J., Sodupe-Ortega, E.,´ Martinez-de Pison, F., 2015. Improving hotel room demand forecasting with a hybrid ga-svr methodology based on skewed data transformation, feature selection and parsimony tuning, in: International Conference on Hybrid Artificial Intelligence Systems, Springer. pp. 632–643.

[43] Veiga, J., Exposito, R.R., Pardo, X.C., Taboada, G.L., Tourifio, J., 2016. Per-´ formance evaluation of big data frameworks for large-scale data analytics, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE. pp. 424–431.

[44] Wanyan, T., Vaid, A., De Freitas, J.K., Somani, S., Miotto, R., Nadkarni, G.N., Azad, A., Ding, Y., Glicksberg, B.S., 2020. Relational learning improves prediction of mortality in covid-19 in the intensive care unit. IEEE Transactions on Big Data .

[45] Wilder-Smith, A., Chiew, C.J., Lee, V.J., 2020. Can we contain the covid-19 outbreak with the same measures as for sars? The lancet infectious diseases 20, e102–e107.

[46] Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang,

W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al., 2020a. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. Journal of thoracic disease 12, 165.

[47] Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al., 2020b. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. Journal of thoracic disease 12, 165.

[48] Ying, S., Li, F., Geng, X., Li, Z., Du, X., Chen, H., Chen, S., Zhang, M., Shao, Z., Wu, Y., et al., 2020. Spread and control of covid-19 in china and their associations with population movement, public health emergency measures, and medical resources. MedRxiv .

[49] Zakary, O., Rachik, M., Elmouki, I., 2016. On the impact of awareness programs in hiv/aids prevention: an sir model with optimal control. Int. J. Comput. Appl 133, 1–6.

[50] Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., Yang, T., Lou, B., Chi, Y., Long, H., et al., 2020. Predicting covid-19 in china using hybrid ai model. IEEE transactions on cybernetics 50, 2891–2904.

[51] Makarand L, M. . (2021). Earlier Detection of Gastric Cancer Using Augmented Deep Learning Techniques in Big Data with Medical Iot (Miot). Research Journal of Computer Systems and Engineering, 2(2), 22:26. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/28

[52] M, T. ., & K, P. . (2023). An Enhanced Expectation Maximization Text Document Clustering Algorithm for E-Content Analysis. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 12–19. https://doi.org/10.17762/ijritcc.v11i1.5982