Prognosticating the effect on Unemployment rate in the post-pandemic India via Time-Series Forecasting and Least Squares Approximation

Ashutosh Agrahari, Ankur Veer, Anshuman Singh, Pawan Singh, *Member, IEEE Computational Intelligence Society*, Baseem Khan, *Member, IEEE*

Abstract—The COVID-19 pandemic adversely influenced and disturbed all the phases of human lives. The number of cases in India has crossed an enormous point of 1.44 millions surpassing Russia and is now third on the list, just behind USA and Brazil. An alarming situation has engulfed whole of the country with many places identified as Containment and Red zones. The search for the cure for this disease is surging high but in due time countries are helpless. They are forced to bow before this pandemic and observe series of lockdowns and social distancing rules. Though these measure are ensuring safety of individuals, they are on the other hand, ruining country's economy. With minimisation in the number of buyers and sellers, India's economy is experiencing a sudden setback which is not easy to get out of. Consequently, it is adversely affecting the employment situation in India. On the brighter side, with students and professionals at their home, job aspirants are getting upteen hours to train themselves. Various investments of foreign companies and industries are undergoing in India currently. With India being the largest hub of young minds, the times post pandemic may be fruitful for the country and may bring bouquets of employment opportunities and thereby help in reinstating the dwindling economy. The current paper aims to analytically visualise these future outcomes that the post-pandemic India might have in store for its citizens. We used time series forecasting on various collected data and combined the statistics of economics-deciding parameters to forecast the trends that might be prevalent in the next 2 years. We also present a thorough analysis of our findings by utilizing paradigms of statistics, economics and machine learning.

Index Terms—Coronavirus, COVID-19, unemployment, time-series forecasting, Prophet, Least squares approximation, vector approximation

1 Introduction

The COVID-19 pandemic has hit the world very harshly. India has now become a hotspot of this pandemic after the USA and Brazil, after surpassing Russia in the number of cases, and is soon to enter stage 3 considering fatality rate. The countries are helpless before this natural phenomenon and are forced to implement lockdowns to prevent its spread. As is evident from many other historical pandemics and epidemics in various parts of the world, there has always been a great deal of adverse effects on the economic conditions of the country or the world as a whole. Consequently, due to lockdowns, trading and working in the industries is stalled which in turn leads to loss of jobs of numerous employees. The employers are then not in the position to support themselves and thereby many companies and industries get shut down. Since industries run a country's economy, and with the industries shutting down or getting in loss, the economy of the entire country gets disrupted. A disease outbreak, being at the scale of the world, affects the entire world. Since most countries are connected through trade and commerce, the pandemic can lead to utter loss of jobs and economy of nations.

The COVID-19 disease has been growing steadily at an exponential rate throughout the world. Since its inception, it has disrupted various sectors of human life including business, health, education and governance. As of July 20, 2020 the number of cases is 14,360,451, out of which 603,285 have lost their life while 8,071,937 have recovered. The economies have been adversely affected and their GDP (Gross Domestic Product) is following a negative slope in today's scenarios. In India, due to sequence of lockdowns, menial laborers have lost their daily-wage jobs, and are left with nothing to sustain in this pandemic-affected economy, and hence many have died not due to the disease but of scarcity of money to support one's family. This is a very arduous situation in the country, and its citizens must not leave hope. Government has been doing its best to control this pandemic. It has taken many initiatives and cut down the cost of COVID-19 tests by half, but still the situation in the country is not under control and is surging high.

The unemployment rate in India had been swinging closely to around 2.7% in previous years but has been exaggerated to an astonishing 15% average during this pandemic, specially between March and May. The rules of social distancing and lockdown has added more to disruption rate of dwindling unemployment rate and consequently on the overall economy of the country. Apart from studying the effects of the pandemic on unemployment rate, inflation rate and economy growth rate, the paper also analyses the conse-

- *A. Agrahari, A. Veer, A. Singh and P. Singh are with the Department of Computer Science and Engineering, Amity University, Lucknow, India. E-mail: ashutoshmathsgenius@gmail.com, ankurveer011@gmail.com, singhanshuman.999@gmail.com, pawansingh51279@gmail.com*
- *B. Khan is with the Department of Electrical and Computer Engineering, Hawassa University, Hawassa P.O. Box 05, Ethiopia. E-mail: baseem.khan04@gmail.com*

Manuscript received Month XX, 20XX; revised Month XX, 202X.

quences of up-skilling at home, on academia and corporate world. We run time-series forecasting on the collected data, and present our findings through comprehensive analysis of the results.

The paper is organized as follows. Section 2 presents a discussion about the works that have been conducted and published in the past. Section 3 outlines the tools and technologies that have been used, and methodology that has been adopted to implement analyses presented in the paper. Section 4 presents the results of the algorithms employed to perform time-series forecasting on the curated dataset. Finally, section 5 summarizes and concludes the paper.

2 Related Work

With the onset of the pandemic, a lot of researches have been conducted by researchers worldwide on the various aspects of the mankind's life from health and hygiene to social and environmental contexts. Sengupta et al. [1] estimated the date when the number of cases would reach the peak and when it will flatten out using methods shared from data analysis and machine learning paradigm. Gupta et al. [2] used time-series forecasting methods like ARIMA to predict the future trends in the rise of the COVID positive cases in India. Poddar and Yadav [3] concluded the relationship between the pandemic and the downfall of Indian Economy based on their null hypothesis. Paul [4] showed the effect of pandemic on Indian economy graphically which included representation of unemployment rate based on CMIE data and various other related parameters. Katris [5] analyzed different machine learning (neural networks, support vector regression) and time series (ARIMA, FARIMA) models to predict unemployment rates of several countries. Karlsson and Javed [6] used both univariate and multivariate time series models like SARIMA, SETAR and VAR, and various macroeconomic variablesto predict unemployment rate and presented the results with 95% forecasting confidence of SARIMA model. The paper also suggested that short term forecasting models give better result than long term.

Most of the researches related to forecasting for COVID have been mainly to predict the number of cases, or predict the unemployment rate using vanilla models and algorithms. Our research focuses on the social sphere of the problems that arose as a consequence of the pandemic. We try to incorporate the sudden spike that happened as a consequence of the pandemic, which made making predictions over Indian economy difficult. In this paper, we handle the shortcomings of the unavailability of reliable data and also the sudden spike that disturbed the underlying mathematics of the models.

3 Methodology Adopted

3.1 Data Collection

There is an unavailability of proper data regarding unemployment rate in India prior to 2016. Adequate data started to be gauged from 2016 onward by CMIE, India. But this much amount of data is not adequate and sufficient for training deep learning and highly computational statistical models - they need good amount of data to learn patterns. So, we also referred to World Bank and the International Labor Organization (ILO) for the data. After data aggregation and cleaning, the data before 2016 was just for the financial years and not for specific months, so we synthetically generated data for months using forward interpolation.

3.2 Heuristic-based Model Training and Prediction

The prediction of future effects has been made using Facebook's time-series forecasting framework Prophet. The accumulated data was first cleaned and pre-processed as per the requirements. Due to the pandemic in 2020 and unavailability of required data, basic implementations of these models could not perform and give reliable results.

To overcome the predilection of the learning models on comparatively larger pre-pandemic historical data and the pandemic-induced sudden hike in unemployment rate, we propose a new computational model for time-series forecasting. We first held 10% of the data, and trained the models on rest of it. We also made two datasets - one with COVID-19 patterns and other without it. The COVID-19 patterns could be seen in Indian economics from the beginning of 2020. Since, we have the actual patterns, we combine the results of models on two datasets and minimize its difference with the actual values by finding the most approximate solution to the set of linear equations that we model. The validation-optimization pipeline could be well illustrated by figure 1.

We aim to minimize the function $aP_1 + bP_2 - C$, where P_1 is the vectors of predictions by the forecasting model that ran on dataset with COVID-19 patterns; P_2 is the vectors of predictions by the forecasting model that ran on dataset without COVID-19 patterns; and C are the actual values for the held timeline. We then converted this equation to be represented using matrices, **AX - C**. Here, **A** is a $2 \times n$ matrix with each column containing P_1 and P_2 respectively, **X** is the vector of parameters a and b that need to be computed, and **C** is the vector of actual values. We then find an approximate solution to this matrix equation using Least Squares Approximation while applying three different underlying models for approximation. Using this proposed technique, we could counteract the issue of sudden noise that got inserted into the data due to the pandemic. After getting optimal values of a and b, we retrain the models on full datasets, and then forecast for the next 24 months timeline by combining the results of two models using the model described in figure 1. Though, the hike is total anomaly in the pattern but it cannot be ignored, so we smoothed down the predicted results of the two models using a heuristic based on the length of timelines for which the two models have been trained on.

4 Results and Discussion

The experimentation and inferencing involves three steps. In the first step, we run the time-series forecasting framework on two types of datasets that we have created. Thereafter, we model the results in the form of matrix equations, and then in the last step, we combine the results and optimizing it by finding suitable parameters a and b using the three different algorithms under Least Squares Approximation, which are -

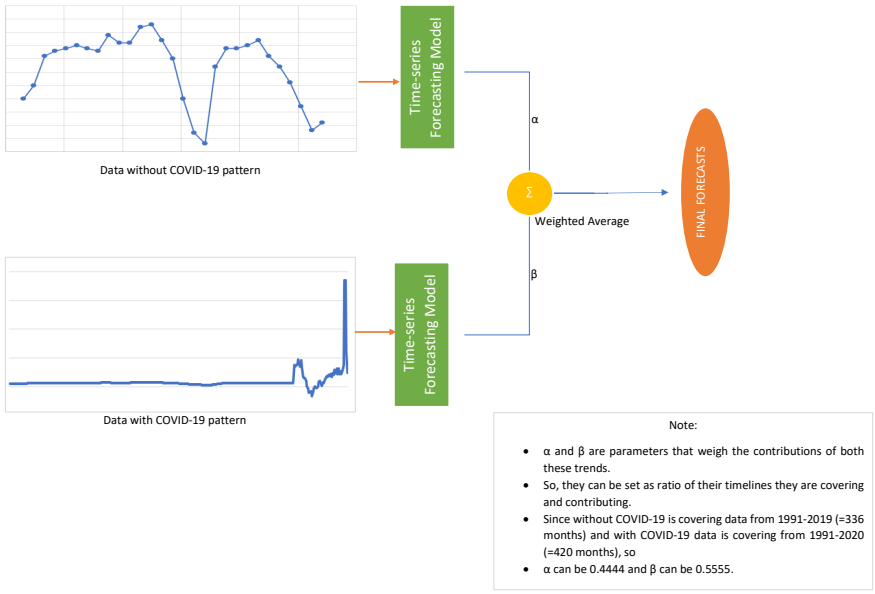

Fig. 1: Computational model for time-series forecasting to counteract the sudden COVID-19 pattern.

1) *Levenberg-Marquardt algorithm*: LMA is an optimization algorithm that combines the best of both the steepest gradient descent method and the Gauss-Newton method. It aims to find the minimum of linear or non-linear function having an array of parameters. The gradient descent method is useful when the initial point is far from minimum point of the function. The Gauss Newton algorithm is useful when initial point is close to the minimum point of the function. The combination of the two helps the algorithm become independent of initial point and hence Levenberg-Marquardt algorithm is derived. It can be expressed by equation 1.

$$\begin{aligned} &x_{n+1} = x_n - (H + \lambda I)^{-1} \bigtriangledown f(x_n) \\ &when, \\ &\lambda \to 0,\ Gauss - Newton's\ Method \\ &\lambda \to \infty,\ Gradient\ Descent\ Method \end{aligned} \tag{1}$$

Here, $f(x)$ is a multi variable non-linear function, H is the Hessian matrix, I is the identity matrix, λ is a hyperparameter, and n is the number of iterations.

2) *Trust Region Reflective algorithm*:
3) *Dogleg algorithm with rectangular trust regions*:

All the training and inferencing procedures were performed on Google's Colab Platform with CPU runtime powered by a single core, two threaded Intel Xeon processor and 12.72 GB of RAM.

4.1 LSTM

LSTM or the Long-Short Term Memory is one of the versions of recurrent neural network architectures that helps to counter-fight the problem of exploding gradients that existed with RNNs.

4.2 ARIMA

ARIMA or the Auto-Regressive Integrated Moving Average is a purely statistical model. It is mainly employed in the time-series models by statisticians and economists. There are two types of models under ARIMA, namely, seasonal and non-seasonal. The non-seasonal model takes into consideration three parameters - p, d and q. Here, p refers to the number of lags to be used to predict the series or it is the order of Auto Regressive term. q refers to the number of lagged errors or it is the order of the Moving Average term. Since, the very first operation involved in ARIMA model is making time series stationary, so, to perform that the most common way is to difference the previous value from the present value. So, here comes the third parameter d which refers to the minimum number of differencing required to make the time series stationary. The non-seasonal model doesn't account for any seasonal patterns, so to add that functionality Seasonal ARIMA or the SARIMA comes into picture. The mathematical form of the model involves two models - Auto Regressive model and Moving Average model. A pure Auto Regressive model can be mathematically expressed by equation 2, where, Y_{t-1} is the lag 1 of time series, β_1 refers to the coefficient of lag 1, ϵ_1 is the error term, and, α refers to the intercept term.

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \epsilon_1 \tag{2}$$

Similarly, Moving Average model can also be depicted mathematically by equation 3, where error terms ϵ_i refer to errors from auto regressive model, ϕ_i are the coefficients of error terms and α is the intercept term.

$$Y_t = \alpha + \sum_{i=1}^{q} \phi_i \epsilon_{t-i} + \epsilon_t \tag{3}$$

Now, the ARIMA model consists of at least one differencing parameter to make time series stationary which combines Auto Regressive model and Moving Average model. The final mathematical form of the ARIMA model can be depicted by equation 4.

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{q} \phi_i \epsilon_{t-i} \tag{4}$$

4.3 Prophet

Prophet is an open-source time-series forecasting model developed by Facebook. The parameters in the model can be easily adjusted for tuning while integrating the core business knowledge. Missing data, outliers can be handled easily and automatically. The three main components of Prophet are *trend*, *seasonality* and *holidays*. The mathematical model underneath Prophet is a decomposable additive model (Eqn. 5) to accommodate the in-dependency of importance and forecasting effects of the components of the model.

$$y_m = \beta_0 + \sum_{n=1}^{t} f_n(x_{mn}) + \epsilon_m \tag{5}$$

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{6}$$

The adapted version of the decomposable additive model can be expressed by the equation 6. Here, $g(t)$ represents trend, $s(t)$ seasonality, $h(t)$ holidays and ϵ_t the error term associated with modelling. The main advantage is that unlike all other models which are natively made for univariate time-series, Prophet very easily accommodates multivariate dimension of time-series data paradigm. Prophet proved to be the most reliable algorithms out of all the three techniques we employed in our experiments. The Prophet architecture is made to train on the univariate data.

4.3.1 Results

As discussed in aforementioned sections, we train a model on two chunks of data, so as to accommodate the sudden hike in the data, which occurred as a consequence of COVID-19, heuristically. Our dataset contains records up till July 2020. Model 1 is trained on dataset from 1991 to 2019, and model 2 is trained on dataset from 1991 to July 2020. The second model takes into account the case of COVID-19 unusual trend in the data. The first model's data is available yearly, so as discussed earlier, we ran forward linear interpolation to generate monthly data.After this, we ran Prophet with seasonality mode as *multiplicative* on both the datasets, and generated predictions for the time-frame ranging from December 2019 to November 2020. Thereafter, we combine the two predictions in three ways to obtain the best forecast pipeline, which not only takes into account the hike in the data but also smoothens out after the effect.

Three parameters were created out of the variables outputted from the least squares approximation algorithm. These parameters as enlisted under Table 1, are generated by generating three different product combinations of the parameters with *data ratio* parameters. This data ratio is the ratio of the timeline length of the two sets of data on which two different models are being trained.

TABLE 1: Parameters used in our heuristic for Prophet model.

Parameter	Var_1	Var_2
Parameter 1	-0.1113	1.1055
Parameter 2	-0.0547	0.5616
Parameter 3	-0.0565	0.5438

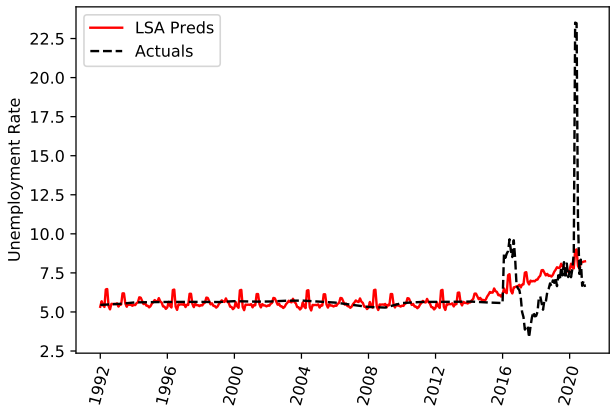

Fig. 2: Predictions on the known timeframe with parameter 1.

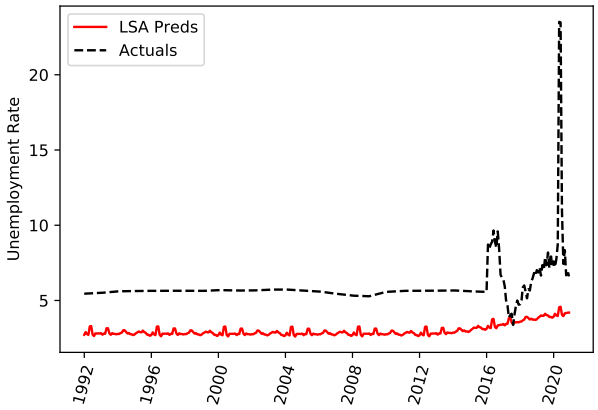

Fig. 3: Predictions on the known timeframe with parameter 2.

Thereafter, predictions from two models are combined as $var_1 \times P1 + var_2 \times P2$, and three different sets of results are obtained. These results are generated as per our heuristic, and we choose the pair of parameters that show the least deviation from the actual data. First we checked the three different models on existing data to see how much they are affected by the sudden hike in the data. These are well elucidated using the plots 2, 3 and 4. Here, we can see that model with vanilla parameters just fits the data, and the other two underfit the data but they gradually pick up the trend and give proper values on the future timeframe.

After checking the pattern followed by the models, we then use our model architecture heuristically combined with

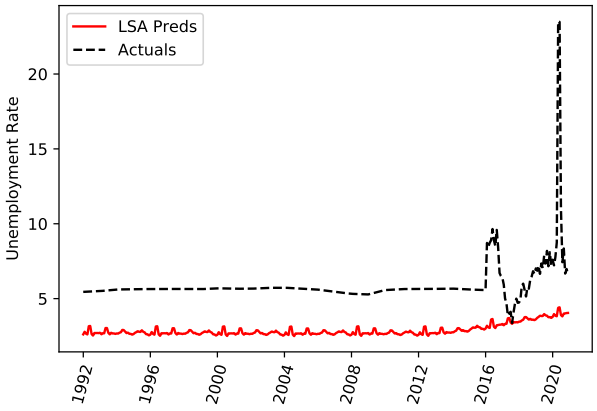

Fig. 4: Predictions on the known timeframe with parameter 3.

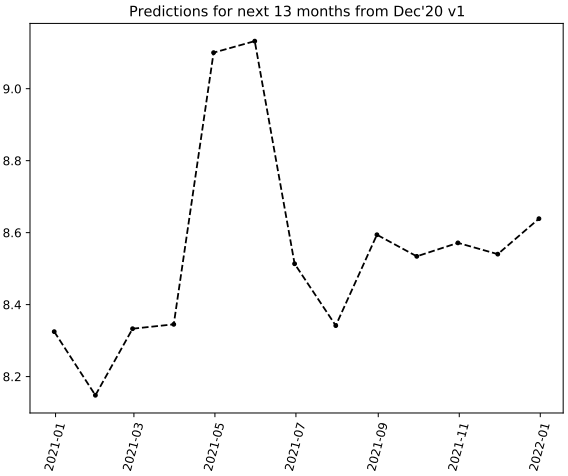

Fig. 5: Predictions on the future timeframe with parameter 1.

three parameters to predict on the future timeframe, that is from December 2020 to December 2021. The results are elucidated by plots 5, 6 and 7. Plot 5 uses vanilla parameters obtained from the Least Squares Approximation algorithm which does not seem to give close to results prevailing in the future timeframe. The other two heuristic's parameters seem to somewhat give significantly good results. After comparing these with the current ongoing data, we conclude that the parameter 3 combination gave the best result.

5 Concluding Remarks

The current study aimed at predicting the conditions that may be prevalent in post-pandemic India after two and five years. The work ran the three best available algorithms for time-series forecasting for the prediction task. The best result obtained out of the three is reported in the paper.

References

[1] Sengupta, Sohini, Sareeta Mugde, and Garima Sharma. "Covid-19 pandemic data analysis and forecasting using machine learning algorithms." medRxiv (2020). doi: https://doi.org/10.1101/2020.06.25.20140004

[2] Gupta, Rajan, and Saibal Kumar Pal. "Trend Analysis and Forecasting of COVID-19 outbreak in India." medRxiv (2020). doi: https://doi.org/10.1101/2020.03.26.20044511

[3] Poddar and Yadav. "Impact of COVID-19 on Indian Economy- A Review" Journal of Humanities and Social Sciences Research, Horizon Journals (2020). doi: https://doi.org/10.37534/bp.jhssr.2020.v2.nS.id1033.p15

[4] Paul, Dhritabrata. (2020). "Covid 19 Impact on Indian economy". doi: 10.13140/RG.2.2.27275.23846.

[5] Katris, C. "Prediction of Unemployment Rates with Time Series and Machine Learning Techniques". Comput Econ 55, 673–706 (2020). https://doi.org/10.1007/s10614-019-09908-9

[6] Meron, D. "Modeling and Forecasting Unemployment Rate in Sweden using various Econometric Measures". Diss. M. SC. Thesis, Örebro University School of Business, Department of Applied Statistics, https://www. diva-portal. org/smash/get/diva2: 949512/FULLTEXT01. pdf., 59-68, 2016.

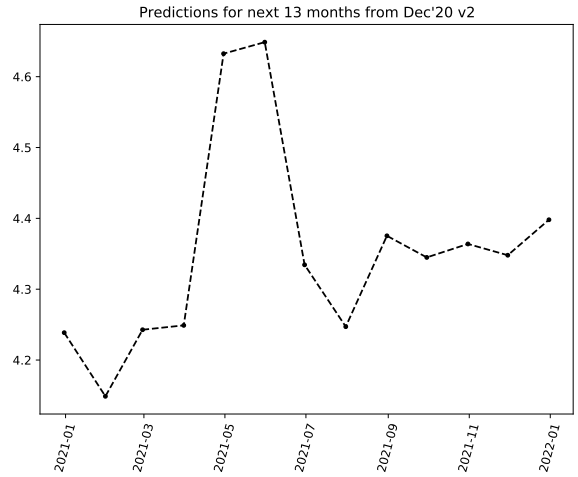

Fig. 6: Predictions on the future timeframe with parameter 2.

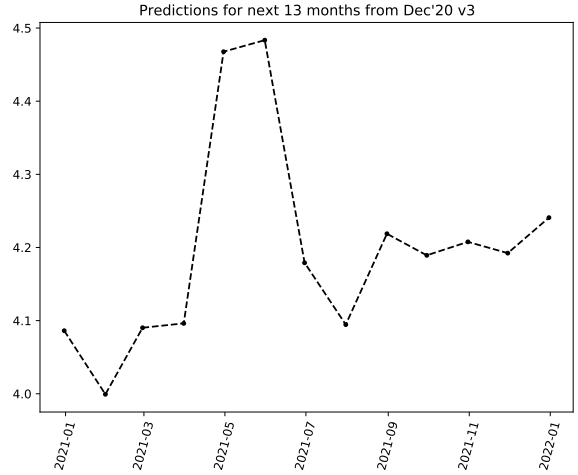

Fig. 7: Predictions on the future timeframe with parameter 3.

Ashutosh Agrahari Ashutosh Agrahari is currently pursuing B.Tech. in Computer Science and Engineering from Amity University, Lucknow, India. His research interests encompass a wide range of fields, including computer vision, algorithm design, optimization techniques, machine learning, deep learning, and data analytics. He also finds interest in working with interdisciplinary departments to work on to tackle real-life problems.

Pawan Singh Pawan Singh received the B.E. degree in Computer Science and Engineering from Chaudhary Charan Singh University, Meerut, India, the M.Tech. degree in Information Technology from Guru Gobind Singh Indraprastha University, New Delhi, India, and the Ph.D. degree in Computer Science from Magadh University, Bodh Gaya, India, in 2013. He is currently Associate Professor in the department of Computer Science and Engineering, Amity University, Lucknow, India. Previously he has been Associate Professor at the School of Informatics, Hawassa Institute of Technology, Hawassa University, Awasa, Ethiopia. He has authored or co-authored number of research papers in the journals of international reputation. His current research interests include software metrics, software testing, software cost estimation, web structure mining, energy aware scheduling, and nature inspired meta-heuristic optimization techniques and its applications.

Baseem Khan Baseem Khan (M'16) received the B.E. degree in electrical engineering from Rajiv Gandhi Technological University in 2008, and the M.Tech. and Ph.D. degrees in electrical engineering from the Maulana Azad National Institute of Technology, India, in 2010 and 2014, respectively. Since 2015, he has been an Assistant Professor with the Hawassa Institute of Technology, Hawassa University, Awasa, Ethiopia. His research interest includes power system restructuring, power system planning, smart grid, meta-heuristic optimization techniques, and renewable energy integration.