International Journal of Big Data and Analytics in Healthcare
Volume 5 • Issue 2 • July-December 2020

COVID-19 Deaths Previsions With Deep Learning Sequence Prediction:
Bacille Calmette-Guérin (BCG) and Tuberculosis Track

Heni Bouhamed, Department of Computer Science, Faculty of Economics and Management, Sfax University, Sfax, Tunisia & Advanced Technologies for Image and Signal Processing Unit (ATISP, Enet'com Sfax), Sfax, Tunisia

ABSTRACT

In this study, the authors use deep learning sequence prediction models for the continuous monitoring of the epidemic while considering the potential impacts of Bacille Calmette-Guérin (BCG) vaccination and tuberculosis (TB) infection rates in populations. Three models were built based on the epidemic data evolution in several countries between the date of their first case and April 1, 2020. The data was based on 14 variables for cases prediction, 15 variables for recoveries prediction, and 16 variables for deaths prediction. Prevision results were very promising, and the suspicions on the BCG vaccination and TB infections rates' implications turned out to be warranted. The model can evolve by continuously updating and enriching data, adding the experiences of all affected countries.

KEYWORDS

BCG, COVID-19, Deep Learning, GRU, LSTM, Monotoring, RNN, Tuberculosis

INTRODUCTION

Many doubts have emerged in recent months about the potential links between BCG vaccination, tuberculosis infection and the spread of COVID-19. Some work (Gupta, 2020; M. Gursel & I. Gursel, 2020; Redelman-Sidi, 2020; Schaaf et al., 2020; Hegarty et al., 2020) has been published in this context without being conclusive. So, to keep digging into the matter is much needed now in order to develop a prediction model to better control the pandemic while trying to know the extent of the three factors' possible implications even though it would still be quite difficult to prove. At the beginning of epidemic, stochastic models were widely used, because a small group of carriers had infected people randomly. After that, researchers turned their attention to deterministic models, which make possible to predict the emergence of infection peaks and to define different control strategies. Work, published in this context, touched on studies on a single country (Xinguang & Bin, 2020; Toshikazu, 2020; Qun et al., 2020; Anzai et al., 2020; Jung et al., 2020), the impact of one or more parameters on the evolution of contagion (Anzai et al., 2020), the comparison between the evolution of the current epidemic with the one of previous versions of the corona virus (McAleer, 2020), the risk estimation of fatal cases (Jung et al., 2020) etc. We suggest moving towards advanced artificial

DOI: 10.4018/IJBDAH.20200701.oa1

intelligence techniques to try and develop a model to predict infection and recovery instances for any selected country given many inputs including BCG vaccination and TB infections rate. Such a model can serve as a reference and a tool to inform the public health professionals, clinicians and decision-makers, enabling them to take coordinative and collaborative efforts to control the pandemic. Models can, also, help to better understand the virus by comparing predictions in different situation.

Several studies invoke the prominence of deep neural networks (DNNs) which surpass the performance of the previous dominant paradigm in diverse machine learning applications (Bouhamed & Ruichek, 2018; Hinton et al., 2012; Mohamed et al., 2012; Ciresan et al., 2010; Yu et al., 2011). Deep Learning is a set of machine learning methods allowing to model data with a high level of abstraction. It is based on articulate architectures of various transformations in the nonlinear space (Bouhamed & Ruichek, 2018; Bengio, 2009). It is also considered as part of the Big Data domain. Current interest for Deep Learning is, not only for its conceptual advances, but also for its technological advances. As a matter of fact, all the available solutions (in terms of models learning) can exploit the immense reservoir of power computing, established through actual modern computers, as well as requesting the main processor (CPU) and the graphic dedicated processors (GPU) (Bouhamed & Ruichek, 2018). A Big Data model can adapt with enormous volume of data and with enormous sequential treatment of numbers exceeding most powerful server capacities (Bouhamed & Ruichek, 2018; Zikopoulos & Eaton, 2011). Since prediction, in our context, depends on observations obtained at previous timings, our scope was more about predicting time sequences. The prediction of recovery cases also depends on the predicted numbers of infected cases, the prediction of deaths also depends on the predicted numbers of infected and recovered cases so the model we are trying to develop must also consider this overlap or dependency of predictions.

Sequence prediction has different problems than other types of supervised learning. The sequence imposes an order to the observations that must be preserved when training models and making predictions. In general, prediction problems that involve sequence data are referred to as sequence prediction problems, although there are several other problems that differ based on the input and output sequences. Many artificial recurrent neural network (RNN) architectures (Brownlee, 2017; Hochreiter & Schmidhuber, 1997) are used in the field of Deep Learning among which Simple Recurrent Neural Network (SRNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Unlike standard feedforward neural networks, RNN has feedback connections. It is capable of not only processing single data points, but also entire sequences of data (such as speech or video). For example, RNN is applicable to tasks such as unsegmented and connected handwriting recognition (Graves et al., 2009), speech recognition (Xiangang & Xihong, 2014) and anomaly detection in network traffic or IDS's (intrusion detection systems). We, then, propose to test this three RNN architectures (SRNN, LSTM, GRU) for this study.

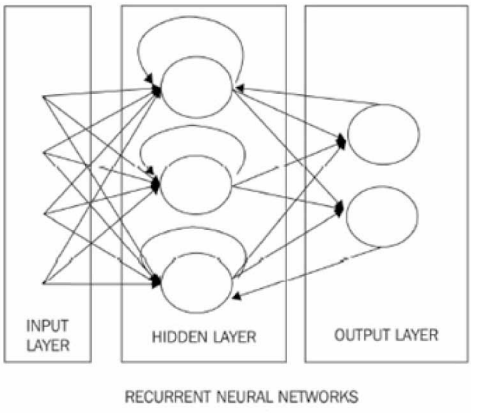
The remainder of this paper is organized as follows: In section 2, we introduce RNN and their three main architectures. In section 3, we treat data, features and methods. In section 4, we proceed with experimentation and discussion. And finally, in the last section, we draw our conclusion and perspectives.

SRNN-LSTM-GRU

A recurrent neural network is a network of artificial neurons with recurrent connections (Figure 1). It consists of interconnected units (neurons) interacting non-linearly and for which there is, at least, one cycle in the structure. The units are connected by weighted arcs (synapses). The output of a neuron is a non-linear combination of its inputs.

SRNN are simple and suitable for input data of various sizes. The training techniques of the network are the same for classical networks, using gradient retro-propagation. However, SRNN may face the problem of gradient disappearance while attempting to memorize too many past events. Some architectures treat this mentioned problem, notably the LSTM and GRU.

Figure 1. RNN



LSTM is the most widely used recurrent neural network architecture in practice, which allows to tackle the gradient disappearance issue. The LSTM network was proposed by Hochreiter and Schmidhuber (1997). The idea behind LSTM is that each computational unit is linked, not only to a hidden state *h*, but also to a state c of the cell that acts as memory. The transition is made by a constant gain transfer equal to 1, so that errors propagate to previous steps (up to 1000 steps in the past) without any gradient disappearance instances (Gers et al., 2002).

GRU were introduced in 2014 by Cho et al. (Cho et al., 2014). Usually when we learn about recurrent networks, we learn about the LSTM first, which could be due to its popularity, or possibly because it was invented first. The GRU is a simpler version of the LSTM. It incorporates many of the same concepts, but it has a much smaller number of parameters, so it can train faster at a constant hidden layer size.

DATA, FEATURES AND METHODS

As a first step, a database has been compiled, representing the evolution of COVID-19 infections for 12 countries from the date of their first cases of infection until March 31, 2020. The data was recovered from ecdc.europa.eu web site and we have chosen countries where the pandemic started early (Table 1).

Table 1. Countries included in our studies

Countries					
China	South Korea	France	Germany	Iran	Spain
Iraq	United K	Italy	Japan	Singapore	Thailand

We have used 13 variables representing demographic, medical and social indicators which data were depicted from the United Nations Development Program-Human Development Report and the World Health Organization. The last variable was the cumulative number of infection cases on day j+1 compared to the first descriptive variable which was the cumulative number of infection cases on day j. The other descriptive variables were as follows:

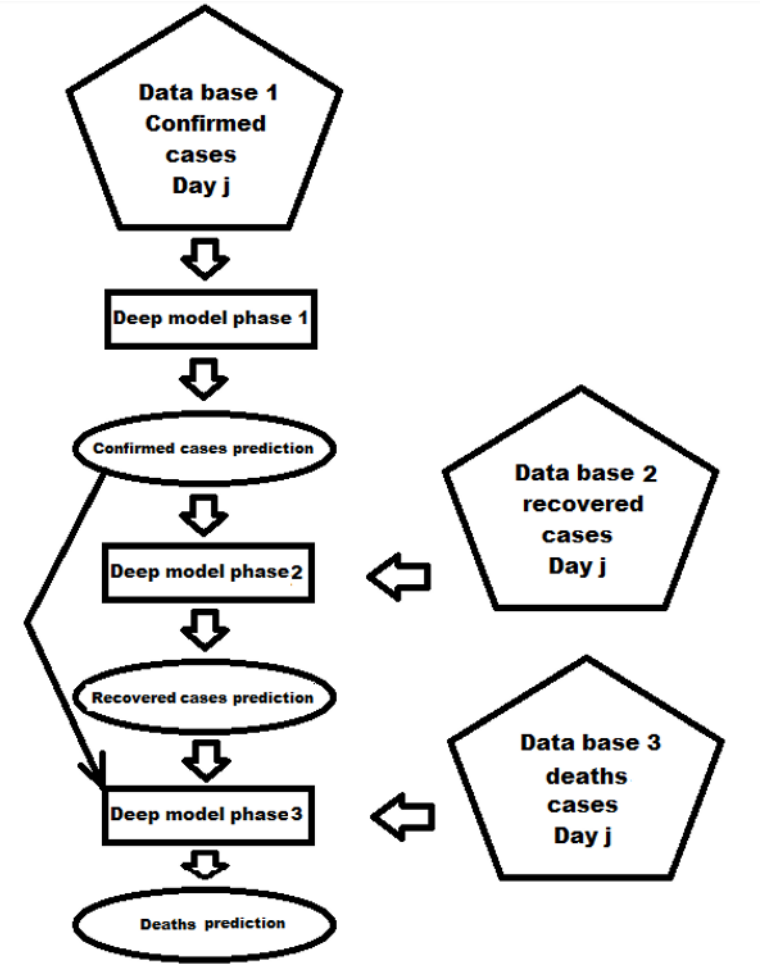
- **BCG Vaccination Policies by Country:** Several studies (Gupta, 2020; M. Gursel & I. Gursel, 2020; Redelman-Sidi, 2020; Schaaf et al., 2020; Hegarty et al., 2020) express doubts about its involvement in the spread of the epidemic;
- **Estimated TB Incidence Rates by Country (2018):** Also several studies (Gupta, 2020; M. Gursel & I. Gursel, 2020; Redelman-Sidi, 2020; Schaaf et al., 2020; Hegarty et al., 2020) share doubts about its implication in the spread of the epidemic;
- **Monthly Average Temperatures:** It turns out that the virus succumbs to high temperatures which could possibly limit contagion;
- **Gross Domestic Product per Capita:** It could give an idea on the financial capacity of the country and its population relevant to fight against the pandemic;
- **Country Population Number:** It was used to control the number of cases according to the number of the population;
- **Country Population Density per Square Meter:** An overpopulated country may be more vulnerable to contagion;
- **Human Development Index:** This index could give an idea on the intellectual and educational capacity of the population, which is positively correlated to prevention measures application such as social distancing;
- **Country Percentage of Health Expenditure:** It gives an idea on the degree of governments involvement in public health, which is crucial in the current fight against COVID-19;
- **Country Number of Visitors per Year:** The number of visitors could be a contributing factor in how fast the epidemic can spread;
- **Country Number of Hospital Beds per 10000 Habitants:** This variable gives an idea on the country's capacity to care for potential patients;
- **Country Number of Days Since the First Case:** This number can help control the epidemic evolution;
- **National Lockdown Establishment:** The lockdown's enforcement could logically limit contagion.

As a second step, a database was set up, representing the evolution of recovered cases for 12 countries (Table 1), starting from the date of their first case of recovery until March 31, 2020. 14 Descriptive variables were used. The last variable was the cumulative number of recovered cases on day j+1 compared to the first descriptive variable which was the cumulative number of recovered cases on day j. For the rest, we kept the same variables used for cases prediction while including the number of cases predicted by the first model and the average age of the population which we considered relevant for the speed of recovery. Note that the datasets of the two phases have a similar distribution.

As a final step, a database was set up, representing the evolution of deaths for 12 countries (Table 1), starting from the date of their first case of recovery until March 31, 2020. 14 Descriptive variables were used. The last variable was the cumulative number of deaths on day j+1 compared to the first descriptive variable which was the cumulative number of deaths on day j. For the rest, we kept the same variables used for recoveries prediction while including the number of cases predicted by the first model phase and the number of recoveries predicted by the second model phase. Note that the datasets of the three phases have a similar distribution.

The training was done in three steps, training a first model for the prediction of the cumulative number of infected cases with database 1, followed by training of a second model for the prediction of the cumulative number of recovered cases with database 2, followed by training of a third model for the cumulative number of deaths with database 3. The prediction process with our third model is shown in Figure 2. A first prediction would be computed based on the phase 1 model followed by a second prediction computed based on the phase 2 model by retrieving the first prediction as input to the second model followed by a third prediction computed based on the phase 3 model by retrieving the first and the second predictions as input

Figure 2. Prevision process



to the third model. We have tested the three most popular RNN architectures: SRNN, LSTM and GRU. The experimentation codes were written with the Python language. TensorFlow and Keras packages were mainly used for the Deep Learning-LSTM and matplotlib package for the visualization of results. All codes, as well as, the database itself is accessible on GitHub (https://github.com/henibouhamed/COVID-19-LSTM). k-fold Cross Validation are used to evaluate the three model phases. The learning process for the three phases was repeated 10 times. For the optimization of the model three phases hyperparameter, we started by testing the use of the layers in an ascending way (one layer then two etc...), cross validation results continuously improved up to 4 layers before stagnating and this is the reason why we adopted four layers for the model parameterization. A dropout was used after each layer to prevent overfitting.

An investigation was launched at the end to try and verify the links between BCG, TB infection and COVID-19. The study involved three countries: France, Germany and Spain. We used our model (in three phases) to predict the number of deaths for these countries according to the following four possibilities:

- **BCG Vaccination:** No, Estimated TB incidence per 100000 population class = 1 (1 (0-9.9),2 (10-99),3 (100-199),4 (200-299),5 (300-499), 6 (>500) see Figure 3 and Figure 4 depicted from World Health Organization report);
- **BCG Vaccination:** Yes, Estimated TB incidence per 100000 population class = 1;
- **BCG Vaccination:** Yes, real Estimated TB incidence per 100000 population class = 2;
- **BCG Vaccination:** Yes, real Estimated TB incidence per 100000 population class = 5.

Figure 3. BCG vaccination in the world: Global Tuberculosis Report by World Health Organization 2019

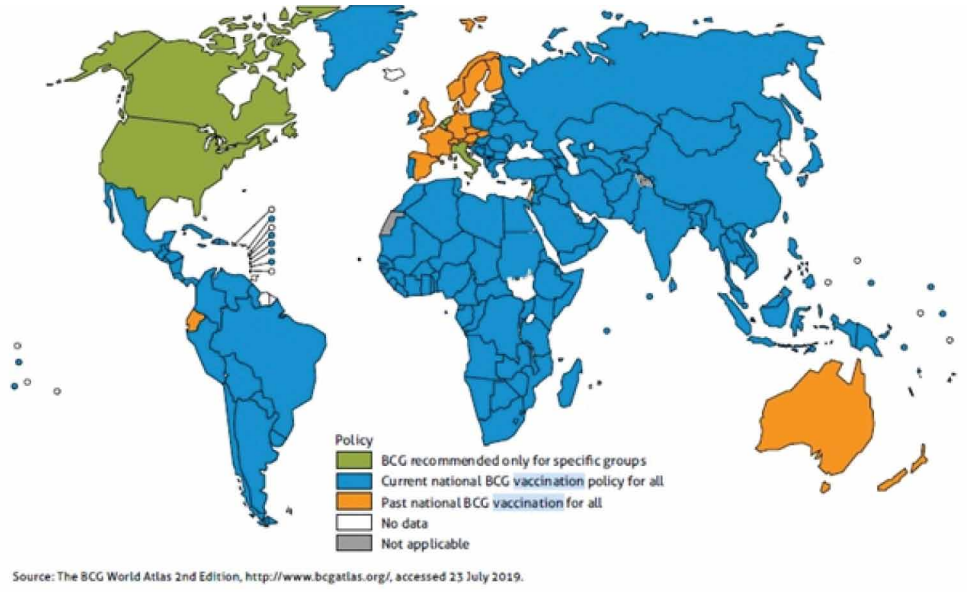
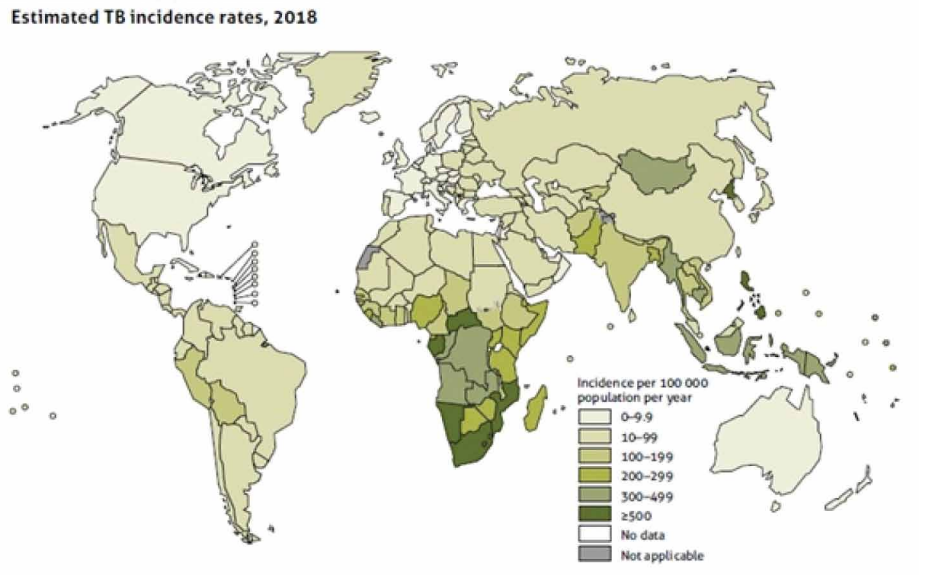


Figure 4. Estimated TB incidence rates: Global Tuberculosis Report by World Health Organization 2019



DISCUSSION

Some studies published in the context of COVID-19 and its possible correlation with BCG and TB mainly presents statistical studies suggesting this possible relationship without presenting a predictive model (M. Gursel & I. Gursel, 2020; Hegarty et al., 2020). Some other studies have simply presented microbiological studies with hypotheses and have not reported decisive conclusion. They also don't have presented a future actions for monitoring and control of the pandemic (Gupta, 2020; Redelman-Sidi, 2020; Schaaf et al., 2020). Outside the context of BCG and TB, several works focused on studying and predicting the pandemic evolution for a single country using deterministic or stochastic models (Xinguang & Bin, 2020; Toshikazu, 2020; Qun et al., 2020; Anzai et al., 2020; McAleer, 2020; Jung et al., 2020). To our knowledge, our work is the only one to propose this conduct in the context of COVID-19, generalized for all affected countries.

Experimental results were expressed in three ways:

- Figures 5, 6, and 7 present the error evolution (loss) for the learning and test data for 10 learnings carried out for the phase 3 with respectively: SRNN, GRU and LSTM;
- Table 2 present the R2 average scores for the phase 1 model with respectively LSTM, GRU, SRNN for 10 learnings. Table 3 present the R2 average scores for the phase 2 model with respectively LSTM, GRU, SRNN for 10 learnings. Table 4 present the R2 average scores for the phase 3 model with respectively LSTM, GRU, SRNN for 10 learnings;
- At the end of this study, we also tried to predict the pandemic deaths evolution (prediction of the evolution deaths by modifying each time the information concerning BCG vaccination and the estimation of TB infections for three countries: France, Germany and Spain (Figure 8, 9 and 10)). According to the average R2 score for each architecture test, the results of LSTM were slightly better (Table 2, 3 and 4), which led us to use it for investigation.

Figure 5. Evolution of the error (loss) for training and test data for 10 learnings carried out for the third phase model training with SRNN

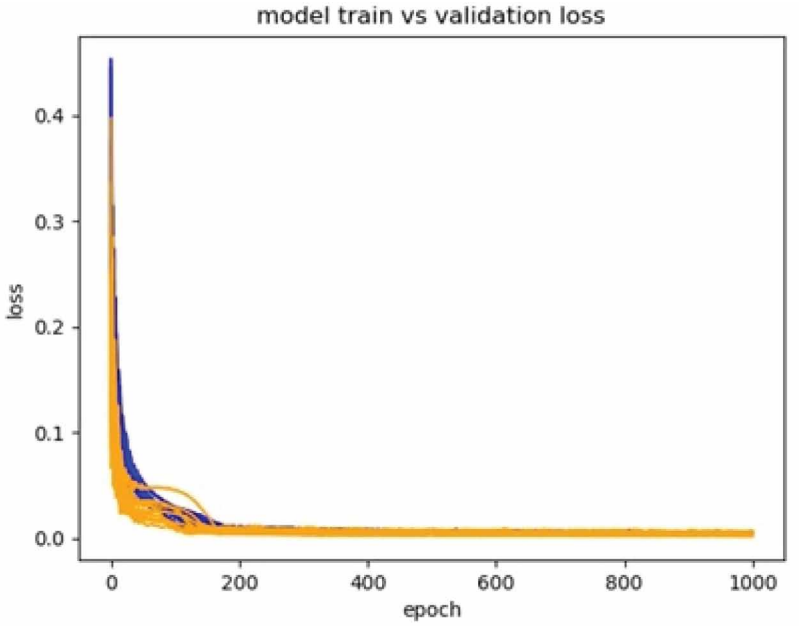
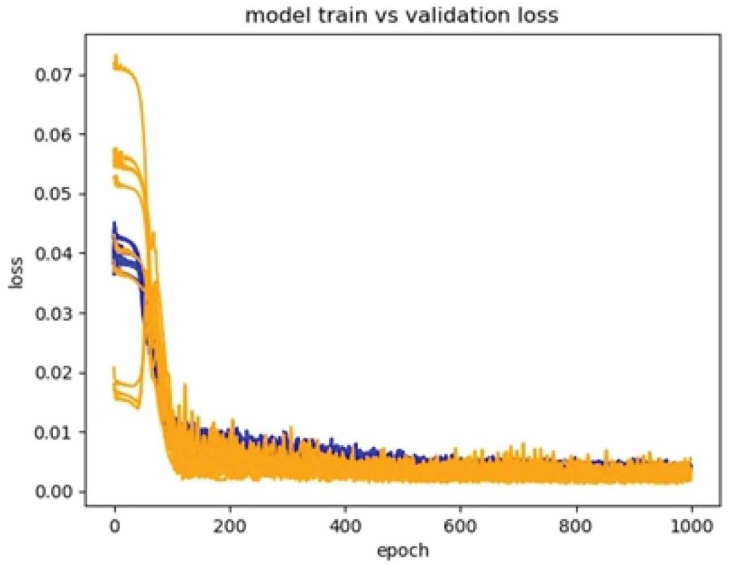


Figure 6. Evolution of the error (loss) for training and test data for 10 learnings carried out for the third phase model training with GRU



Figures 5,6 and 7 show the evolution of the error (loss) during learning with the learning data (in blue) and the test data (in yellow) concerning the third model (death prediction). The prediction error stabilizes close to zero after 500 epochs. These results as well as the R2 scores found (Table 2, 3 and 4) allowed us to conclude that the models' results were globally reliable according to the three RNN architectures tested. The LSTM was slightly better but learning the two other architectures was

Figure 7. Evolution of the error (loss) for training and test data for 10 learnings carried out for the third phase model training with LSTM

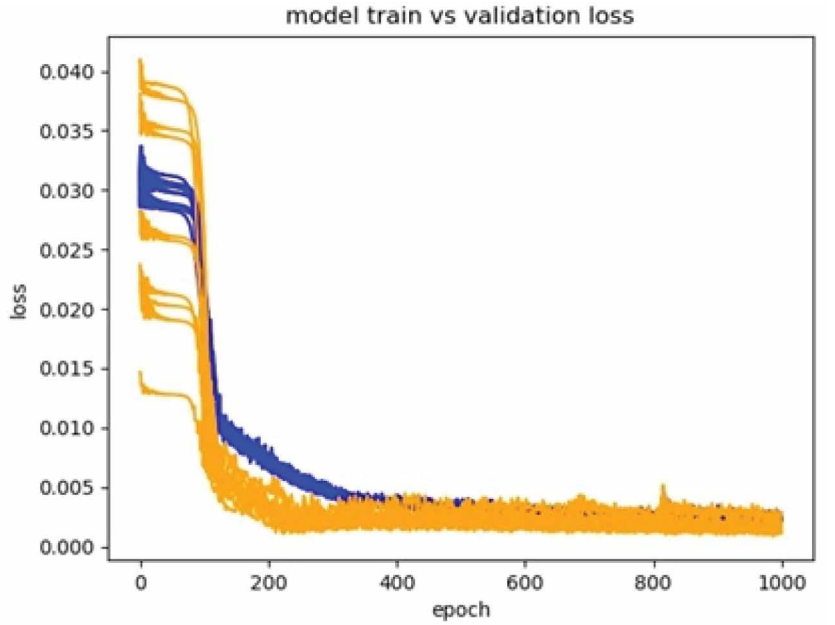


Table 2. R2 average scores for the first phase model for 10 iterations of training

Architecture	Average
LSTM R2	0.999
GRU R2	0.973
SRNN R2	0.959

Table 3. R2 average scores for the second phase model for 10 iterations of training

Architecture	Average
LSTM R2	0.996
GRU R2	0.964
SRNN R2	0.952

Table 4. R2 average scores for the third phase model for 10 iterations of training

Architecture	Average
LSTM R2	0.976
GRU R2	0.944
SRNN R2	0.912

faster. The results of the third phase (deaths predictions) were very encouraging even though the study involved only 12 countries until March 31, a date until which experiences on deaths evolution were lacking.

In the final phase of the studies, Figures 8, 9 and 10 show the predicted deaths for Spain, France and Germany during the first 6 days of April according to the different scenarios concerning bcg vaccination and incidence of tuberculosis in the population. The involvement of the BCG vaccination and the percentage of TB infections, revealed some surprising trends. BCG vaccination significantly decreased the evolution of deaths, also the increase in the percentage of TB infections have a considerable effect in this third phase of the study. Nevertheless, it is very difficult to draw radical conclusions in one direction or the other, and our study may eventually inspire the scientific community to continue research on the matter.

We propose, as a perspective to this work, first, to add the data of all other countries affected by the virus, then to create an automatic process to update the results. This model may help control the pandemic and may help in making the right decisions in advance.

CONCLUSION

In this study, we have used a Deep Learning sequence prediction models (SRNN, LSTM and GRU) for the continuous monitoring of the infection, recovering and deaths processes making investigations about the impact of Bacille Calmette-Guérin (BCG) vaccination and the incidence of tuberculosis (TB) infection rate in populations. Models were built based on the epidemic data evolution of several countries between the date of their first case and March 31, 2020. The data was based on 14 variables for case prediction, 15 variables for recoveries prediction and 16 variables for deaths prediction.

Figure 8. Prevision of deaths number for France

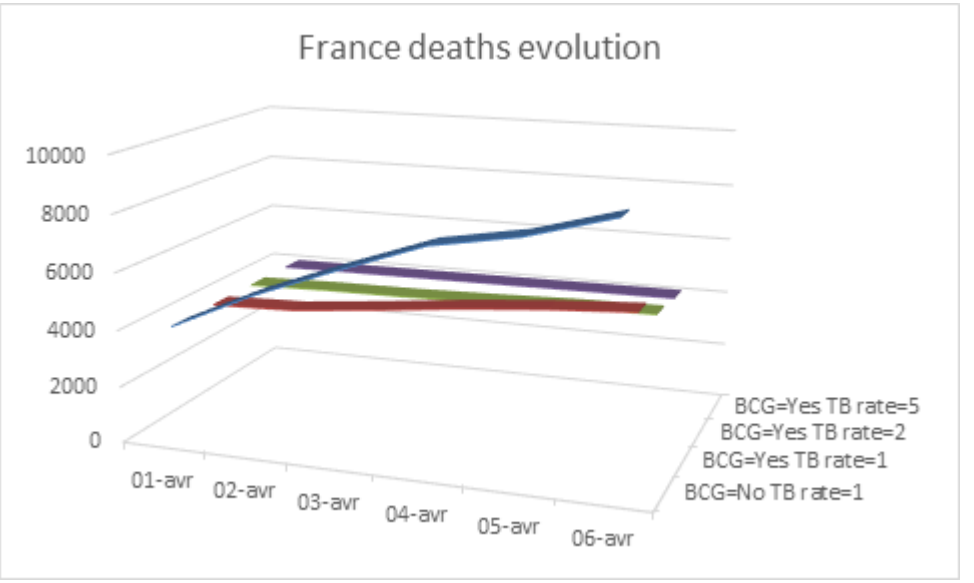
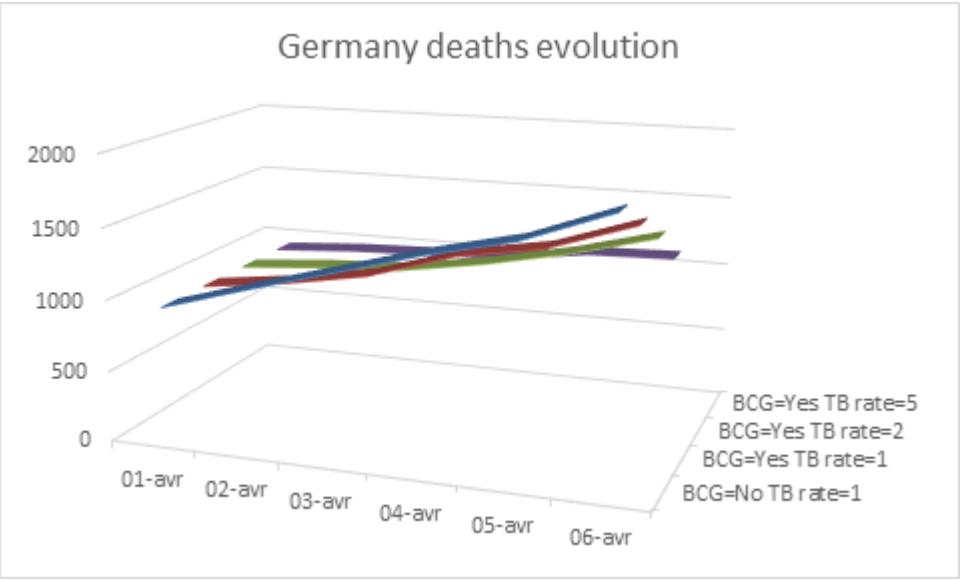


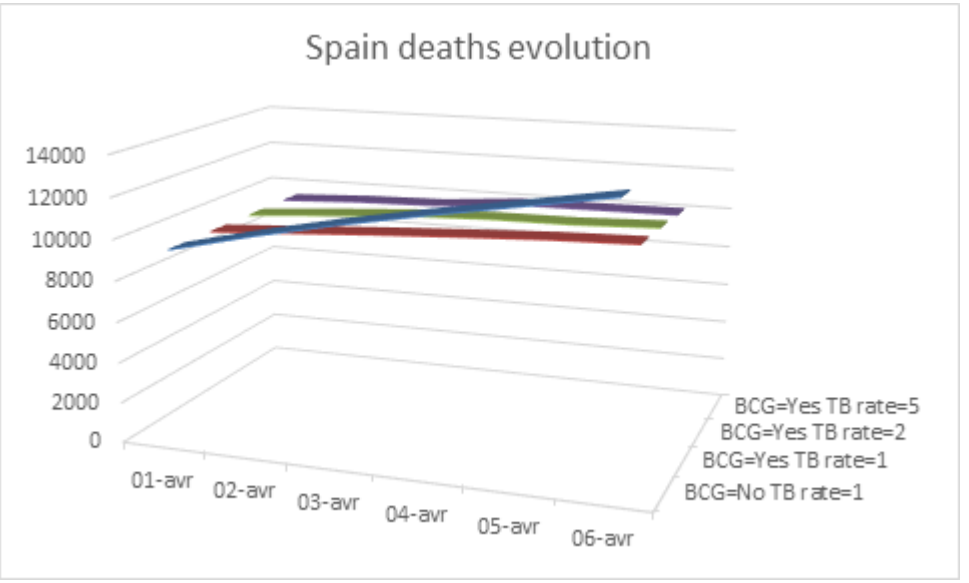
Figure 9. Prevision of cases number for Germany



Prevision results were very promising and the intuition behind the BCG vaccination and TB infections rate implications turned out to be quite relevant. The model can evolve by continuously updating and enriching data by adding experiences of all affected countries.

The methods in this study can help inform public health professionals, clinicians and decision-makers to take coordinative and collaborative efforts to control the epidemic.

Figure 10. Prevision of deaths number for Spain



The perspective for this work is, first, to add data from all other countries affected by the virus, then to create an automatic process to update the results. This model may help control the pandemic and may help making the right decisions in advance.

REFERENCES

Anzai, A., Kobayashi, T., Linton, N. M., Kinoshita, R., Hayashi, K., Suzuki, A., Yang, Y., Jung, S., Miyama, T., Akhmetzhanov, A. R., & Nishiura, H. (2020). Assessing the Impact of Reduced Travel on Exportation Dynamics of Novel Coronavirus Infection (COVID-19). *Journal of Clinical Medicine*, *9*(601), 601. Advance online publication. doi:10.3390/jcm9020601 PMID:32102279

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning, 2*(1), 1-127.

Bouhamed, H., & Ruichek, Y. (2018). Deep feedforward neural network learning using Local Binary Patterns histograms for outdoor object categorization. *Advances In Modelling And Analyses B*, *61*(3), 158–162. doi:10.18280/ama_b.610309

Brownlee, J. (2017). *Long Short-Term Memory Networks With Python*. Machine Learning Mastery Edition.

Cho, K., Van, M. B., Gulcehre, C., Bougares, F., Schwenk, H., & Ben-gio, Y. (2014). *Learning phrase representations using rnn encoder-decoder for statistical machine translation.* arXiv:1406.1078

Ciresan, D., Meier, U., Gambardella, L., & Schmidhuber, J. (2010). *Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition.* CoRR abs/1003.0358

Gers, F., Schraudolph, N., Schmidhuber, J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research, 3*, 115-143.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868. doi:10.1109/TPAMI.2008.137 PMID:19299860

Gupta, A. (2020). Is Immuno-modulation the Key to COVID-19 Pandemic? *Indian Journal of Orthopaedics*, *54*, 394–397. PMID:32341599

Gursel, M., & Gursel, I. (2020). Is Global BCG Vaccination Coverage Relevant To The Progression Of SARS-CoV-2 Pandemic? *Medical Hypotheses*, 109707. Advance online publication. doi:10.1016/j.mehy.2020.109707

Hegarty, P., Kamat, A., Zafirakis, H., & Dinardo, A. (2020). *BCG vaccination may be protective against Covid-19*. VOX CEPR Policy Portal. doi:10.13140/RG.2.2.35948.10880

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. doi:10.1109/MSP.2012.2205597

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735 PMID:9377276

Jung, S., Akhmetzhanov, A. R., Hayashi, K., Linton, N. M., Yang, Y., Yuan, B., Kobayashi, T., Kinoshita, R., & Nishiura, H. (2020). Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. *Journal of Clinical Medicine*, *9*(523), 523. Advance online publication. doi:10.3390/jcm9020523 PMID:32075152

McAleer, M. (2020). Prevention Is Better Than the Cure: Risk Management of COVID-19. *Journal of Risk Financial Managment*, *13*(46), 46. Advance online publication. doi:10.3390/jrfm13030046

Mohamed, A., Dahl, G., & Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14–22. doi:10.1109/TASL.2011.2109382

Qun, L. Med, M., Xuhua, G., Peng, W., Xiaoye, W. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *The New England Journal of Medicine*. Advance online publication. doi:10.1056/NEJMoa20013

Redelman-Sidi, G. (2020). Could BCG be used to protect against COVID-19? *Nature Reviews. Urology*, *17*(6), 316–317. Advance online publication. doi:10.1038/s41585-020-0325-9 PMID:32341531

Schaaf, H. S., Preez, K., Kruger, M., Solomons, R., Taljaard, J. J., Rabie, H., Seddon, J. A., Cotton, M. F., Tebruegge, M., Curtis, N., & Hesseling, A. C. (2020). Bacille Calmette-Guérin (BCG) vaccine and the COVID-19 pandemic: Responsible stewardship is needed. *The International Journal of Tuberculosis and Lung Disease*. Advance online publication. doi:10.5588/ijtld.20.0267

Toshikazu, K. (2020). Prediction of the Epidemic Peak of Coronavirus Disease in Japan 2020. *Journal of Clinical Medicine, 9*(789).

Xiangang, L., & Xihong, W. (2014). *Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition.* arXiv:1410.4281

Xinguang, C., & Bin, Y. (2020). *First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: realtime surveillance and evaluation with a second derivative model*. Global Health Research and Policy. doi:10.1186/s41256-020-00137-4

Yu, D., & Deng, L. (2011). Deep Learning and its applications to signal and information processing (exploratory DSP). *IEEE Signal Processing Magazine*, *28*(1), 145–154. doi:10.1109/MSP.2010.939038

Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

Heni Bouhamed received Master of Computer Research, Engineering, and Information Systems (IGIS) from University of Rouen, France in 2006 and Ph.D. degree in Computer Systems and Engineering at the National School of Engineers of Sfax, under joint supervision with the University of Rouen in France, in 2013. He is currently working as Senior Lecturer in Department of Computer Science, University of Sfax, Tunisia since 2015. He is a member of Advanced Technologies for Image and Signal Processing unit (ATISP, Enet'com Sfax) and Co-Founder of DataCamp-Training & Consulting. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences. His main research work focuses on Big Data Analytics and Deep Learning. He has 14 years of teaching experience and 13 years of Research Experience.