eVision: Forecasting the spread of Tuberculosis in India with Deep Learning

Juan Zuluaga, Michael Castillo, Divya Syal, Andres Calle, and Navid Shaghaghi
Ethical, Pragmatic, & Intelligent Computing (EPIC) Laboratory
in collaboration with the Healthcare Innovation & Design (HID) Program
Departments of Bioengineering (BIOE), Computer Science & Engineering (CSEN),
Information Systems & Analytics (ISA), and Mathematics & Computer Science (MCS)
Santa Clara University, Santa Clara, California, USA
{jzuluaga, mvcastillo, dsyal, acalle, nshaghaghi}@scu.edu

***Abstract*—Humanity has battled tuberculosis for all of recorded history. Some studies estimate that Mycobacterium tuberculosis may have been around as long as 3 million years but it was only in 1834 when Johann Schonlein officially presented the characteristics of it. Even though the TB epidemic has touched all corners of the world, Africa and Asia are the regions that currently suffer the worst consequences. The purpose of this study is to construct a model within the eVision forecasting environment, capable of forecasting the trend of Tuberculosis cases in India, as India is the country that accounts for the largest percentage of TB cases and deaths worldwide. And being able to make predictions for India may also lead to successful results for other regions in Asia and Africa.**

In order to do so, this study presents different test cases that show the effectiveness of the model, varying the number of steps for each one of the data sets created. It's important to note, that these data sets are combinations of data gathered from the states with the most TB cases in India in the last years, as well as the total data for India, and supplemental data from Google Trends, as a way to facilitate the machine learning model. Even though the final results were respectable compared to past research done on India and other countries, the model nevertheless has a limitation on the number of weeks ahead which the predictions are still considered to be good; with 7 weeks being the optimal result.

***Index Terms*—Deep Learning, eVision, Forecasting Spread of Disease in India, Long short-term memory (LSTM), Outlier Correction, Tuberculosis (TB)**

I. Introduction

Tuberculosis (TB) is a contagious disease that by many accounts infected Approximately one third of the world's population by the year 1990, and continues to cause the death of millions per year till this day [1]. Eradicating this epidemic has been one of the toughest challenges for the international community in recent history, partly because of the racial and ethnic disparities of the victims of this disease. Even in the modern era, Tuberculosis maintains a high mortality rate in the populations where it stubbornly persists. An estimated 1.5 million people were victims of this pandemic during 2020, placing TB as the second leading infectious killer, only behind COVID-19 [2]. Since forecasting the number of cases several weeks in advance may be a critical step in stopping the spread of TB, the use of data science and machine learning seem invaluable.

Initially, the focus of this research was centered on the spread of TB in the US. It was determined that the most regular spread of the disease in the US occurred within US prisons. But the data revealed that while there are a few cases of TB arising in US prisons regularly, the numbers have dropped significantly over the last ten years and followed a relatively flat trend. However, the larger obstacle was that the data is protected so viewers cannot publicly access state-level TB cases in a year by year format, but rather five years at a time, which is not granular enough for meaningful time series forecasting usable for weekly predictions. The focus thus shifted to investing where TB was and still is relatively abundant: Asia. India was chosen for this study as the country includes the highest reported amount of TB worldwide and because it includes both highly developed as well as rural regions grappling with the disease. During the year 2020, 86% of the total number of TB cases were accounted for in 30 countries, with India leading the list in total cases, followed by China, Indonesia, Bangladesh, The Philippines, Pakistan, Nigeria and South Africa [2]. Specifically, India accounts for approximately 2.6 million of the 10 million cases worldwide in 2021, with an incidence rate around 188 cases per 100,000 population [3]. This paper mainly focuses on tuberculosis in India and the different regions that are in significant danger of a health crisis due to TB outbreaks.

Using India as the first case-study provides the advantage of a vast data set to train the model with, since the data for other countries such as South Africa or China is not nearly as expansive and available as India's. Over the last 30 years, China has significantly reduced TB incidence and mortality, decreasing incidence rates by 24 percent from 2010 to 2019. This success parallels further restrictions on the public availability of Chinese TB data, as the Chinese government loses interest in the health problem year by year [4].

Of further interest, the Indian sub-continent has recently been affected by COVID-19 in addition to the ongoing TB epidemic [5]. Studying the ramifications of the combination of these pandemics is an important field of study itself. The overtaxing of hospitals during the height of the pandemic along with deferral of patients seeking care due to fear of infection would have clearly affected spread patterns.

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00333

Much research has been done concerning the abundance of TB in India, with the delay in diagnosis plus a lack of adequate treatment being underlying factors of the high current TB burden. Delayed diagnosis of TB can enhance the transmission of infection, worsen the disease, increase the risk of death, and may even be a reason for why TB incidence has not substantially declined despite the global scale-up of DOTS (Directly Observed Treatment Short course) strategy in which TB patients are observed swallowing each dose of their medicines by a health worker or trained volunteer [6].

The main contributions of this paper are reporting on the:

- creation of a Long short-term memory model to forecast ahead of time the number of TB cases in India and its largest regions. Since India is the epicenter of this epidemic, if the model works successfully for India, replication of this is conceivable for countries such as China, Indonesia and Bangladesh.
- exploration of a comparison between the number of cases reported and the number of searches on the Google search engine for the term "Tuberculosis".
- finding a correlation between the significant decrease in the number of TB cases during 2021 and the Coronavirus wave that took place in the country during the second quarter of that year.

This paper is organized as follows: Section II describes the background and related work. Sections III and IV respectively describe the phases of the methodology and the different results obtained using the model, as well as trying to find an explanation for different peculiarities discovered in the data. Sections V describes the future steps for this work as well as the aspects that can be improved, and section VI provides some concluding remarks.

II. Background and Related Work

Throughout the years much research on how to forecast the spread of tuberculosis in advance has been conducted, especially in India. Different models and techniques have been developed for the purpose of being prepared for the different TB waves that affect India annually. Each of the papers discussed below presents different characteristics and achieve different goals, using in many cases data sets with complex numerical distribution and outliers. This section presents a review on some of the models developed in older studies, the characteristics of the data sets utilized in them, and their findings.

A. Seasonality of Tuberculosis in Delhi, India: A Time Series Analysis [7]

This study aimed to discover the regular variation of tuberculosis in Delhi, India depending on the season. The methodology employed in this research was a retrospective record based study of a Directly Observed Treatment Short course (DOTS) centre located in the south district of Delhi where Six-years worth of data from January 2007 to December 2012 was analyzed. It's important to note that in other parts of the world a consistent pattern for this variation has not been fully defined, making this one of the first studies to do so, albeit with approximately 70 percent variability when using Winter's multiplicative model. This study was conducted exclusively for Delhi, since the health centre the authors obtained the data from services this specific region.

B. Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network [8]

For the prevention and management of TB, reliable forecasting is helpful. Therefore, this study proposes an approach to predict tuberculosis outbreaks, which is a significant public health issue in China. In order to predict the incidence of TB from January 2007 to March 2016, this study suggests a hybrid model that combines an autoregressive integrated moving average (ARIMA) with a nonlinear autoregressive (NAR) neural network. It was compared how well the hybrid model and ARIMA model performed at making predictions. The paper mainly focuses on China as an epicenter of TB. However, not including more data points from different Chinese provinces might be a limiting factor for the final output. Furthermore, this study is scraping the data in a monthly basis, which has advantages in gathering a more concise data set but weekly scraping would have enabled quicker (weekly rather than monthly) predictions.

C. Seasonality and Trend Forecasting of Tuberculosis Prevalence Data in Eastern Cape, South Africa, Using a Hybrid Model [9]

In this study, the seasonality of TB incidence in South Africa is examined and compared with the prediction abilities of the seasonal autoregressive integrated moving average (SARIMA) and neural network auto-regression (SARIMA-NNAR) models. As for the methodology employed, the data on TB incidence cases were extracted from the Eastern Cape Health institution report of the electronic Tuberculosis Register between January 2010 and December 2015. The TB data from 2010 to 2015 was analyzed and tested using a SARIMA model as well as a SARIMA-NNAR model, which combines a SARIMA model and a neural network auto-regression model. Mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean percent error (MPE), mean absolute scaled error (MASE), and mean absolute percentage error (MAPE) simulation performance parameters were all used in order to evaluate how accurate the model was performing for different cases.

D. Seasonality of tuberculosis in rural West Bengal: A time series analysis [10]

This investigation was done to create a univariate time series model to evaluate the seasonality of tuberculosis in rural areas of West Bengal. A total of 1507 new cases of tuberculosis that were recorded in the TU's TB register between January 2008 and December 2011 were used for this investigation. As for the methodology employed, the SPSS 16.0 version was used to apply the seasonal adjusted factor (SAF), autocorrelation function (ACF), partial autocorrelation function (PACF), and

seasonal autoregressive integrated moving average (SARIMA) procedures. The difference between this investigation and the one in the present paper is the context itself. The investigation located in West Bengal is looking to construct a model mostly suited to predict for rural areas, since the format of the information and characteristics differ considerably from the urban data sets employed by eVision's model.

III. eVision's TB Forecaster

In this section some of the details, characteristics, and strategies introduced for this research are explained. Particular focus will be on the selection of features added to the model in order to improve performance.

A. Data Gathering

The most granular data source for the spread of TB in India was from the Indian government – called "Nikshay" – which includes detailed data on a daily basis since 2017 and also separates private vs. publicly reported cases. However, with "Nikshay" there was no mechanism to download the state-level data at any time interval in total since 2017, so the raw weekly and state-based TB case data had to be downloaded for each day from 2017, one at a time. Then, a script was utilized to aggregate all the data into one data set. In addition to combining and parsing the data, it had to be cleaned as there were a certain few Indian territories – i.e. Lakshadweep in the Arabian Sea – that reported a few TB cases very irregularly and therefore the province was occasionally included in one of the weekly and state-based TB case data sets. Given this inconsistency, each data set had to be generalized post-download, then aggregated. Based on this raw data set, several other data sets were created, of which the most important include:

- the corresponding normalized data set using z-scores to account for the large range of TB cases;
- the raw data for Google searches of the disease "Tuberculosis" over the same sample date range and format by Indian states, using Google Trends; and
- the normalized data for Google searches of the disease "Tuberculosis" over the same sample date range and format by Indian states, using Google Trends.

The resulting data set presents a week by week report of the number of Tuberculosis in all the regions of India, from 2017 to the start of 2022. For a deeper analysis, some sub data sets of the regions in India most affected by TB were also obtained. These regions were specially critical because in some of the weeks they reached a peak of 50000 cases reported.

B. Modeling

The decision to using an LSTM as the model was based on the influenza and COVID-19 forecasters for the United States and select other countries from previous work done on the eVision tool by the EPIC lab and Healthcare Innovation and Design Program [11] [12] [13] [14] [15]. All three diseases have similarities in data distribution, and all three studies share similar structure. For instance, the addition of different features such as Google trend and state by state data were tested in order to enhance the performance of the model [16] [14].

In addition, Long short-term Memory (LSTM) neural networks provide plenty of advantages here. The cells can learn more parameters, making them even more powerful additions to long-term memory. This makes LSTMs the most powerful Recurrent Neural Network for prediction, especially when there is a long-term trend in the data [17] which is the case for Influenza, COVID-19, and TB data. The LSTM was trained over the course of 200 epochs, with a batch size of 250.The model was tested with different learning rates to improve its accuracy. 0.1 for initial learning rate performed best overall, in order to avoid data loss. The LSTM layer was provided with a sequence input layer with a vector of size 5 and a fully connected layer of size 1.

IV. Results

As can be seen in figure 1, initially the results were not as good as expected.In order to find the best possible result, different combinations of data sets were tested.

As aforementioned, the data set extracted from the official Indian government was originally reported state by state. As has been seen in previous research on influenza in the United Sates there are numerous relationships to explore between national influenza levels and state influenza levels [13]. Finding leading states in the pandemic can provide a significant boost in accuracy to the national model. For instance, one possibility examined is training the model using the states with the highest total levels of TB cases, in this case the highest being Uttar Pradesh, as features supporting the national model. Figures 2 and 3 respectively show the incorporation of only Uttar Pradesh data and All state data as additional parameters.

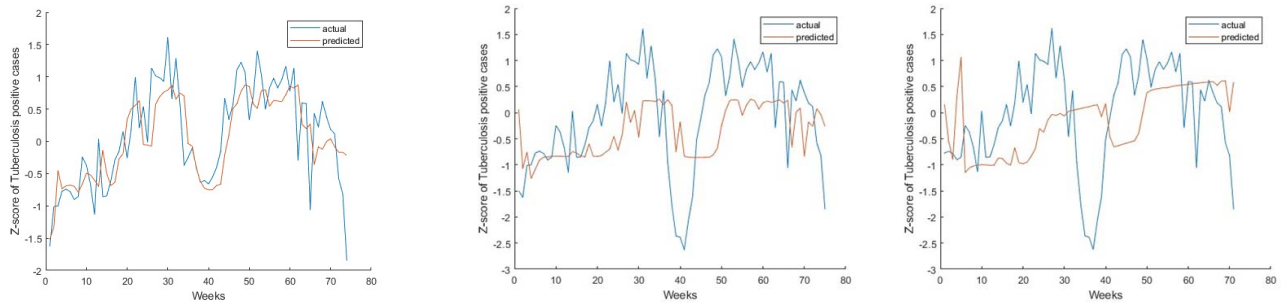
Another way to strengthen the accuracy of the model was to incorporate Google Trends data, which captures the level of interest around a specified term during some subset of time on Google Search. In this study, the term "Tuberculosis" was chosen for obtaining relevant data and the results can be seen in Figure 4.

Note that the y-axis in the graphs represents the Z-score of Tuberculosis positive cases. Z-score is a measure of how many standard deviations below or above the population mean, a raw data point is [18].

A. Results Analysis

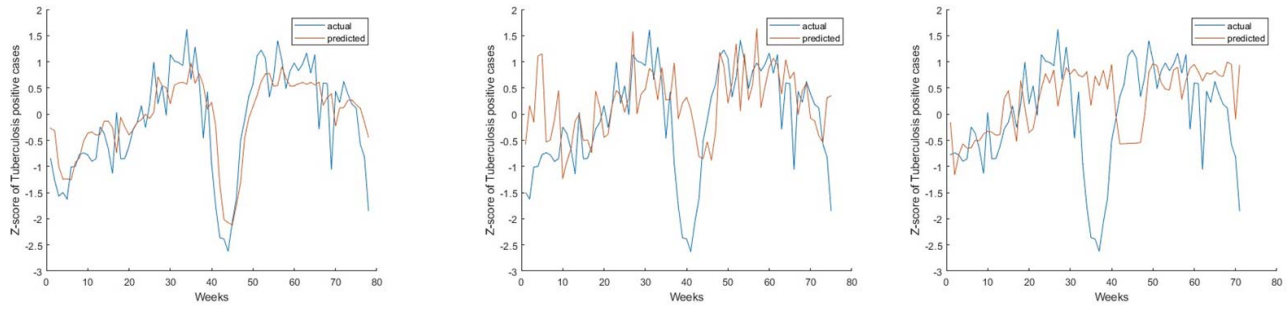
In order to validate the model, Mean Absolute Percentage Error (MAPE) was applied, which is considered to be the ideal measure to test time series forecasting models. Table I presents the MAPE for each one of the test cases presented. This number in conjunction with the graphs gives a better representation of how good the model is performing.

The actual and predicted results are not more than 2 standard deviations apart, making the gap between these 2 not as big as it might seem. However, as mentioned in the methodology, between weeks 37 and 50 an enormous drop in the data reported is visible, which might be the cause for the subpar results. The datapoints on this graph's section relate to April



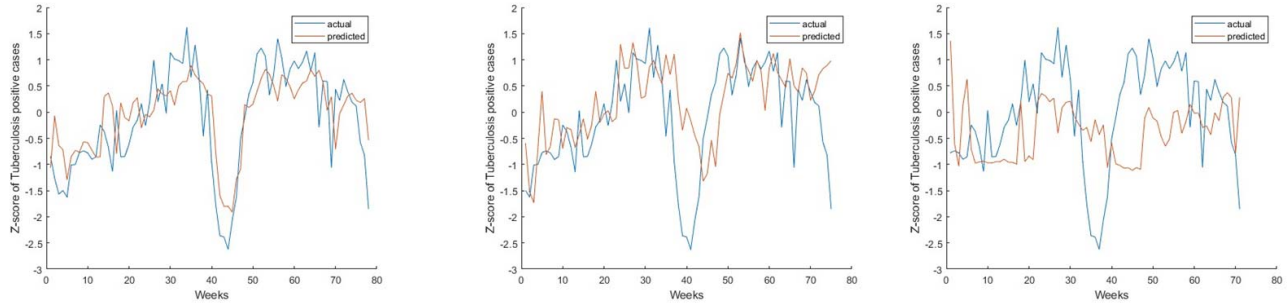
(a) India 1 week ahead (b) India 4 weeks ahead (c) India 8 weeks ahead

Fig. 1: Forecasting using India National Data Only



(a) India + Uttar 1 week ahead (b) India + Uttar 4 weeks ahead (c) India + Uttar 8 weeks ahead

Fig. 2: Forecasting using India National Data + India's State of Uttar Data as additional parameters



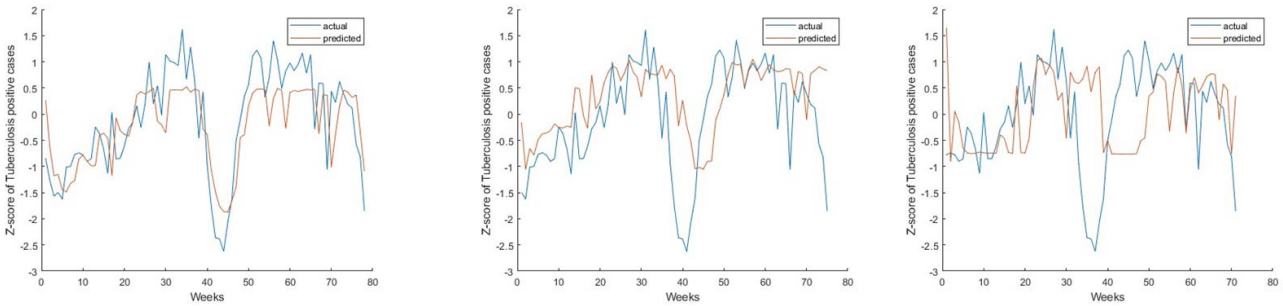
(a) India + All States 1 week ahead (b) India + All States 4 weeks ahead (c) India + All States 8 weeks ahead

Fig. 3: Forecasting using India National Data + India's All States' Data as additional parameters

TABLE I: Accuracy by data set used

data set Entry / # of weeks ahead	1	4	8
India	6.03	8.83	11.22
India + Uttar Pradesh	5.21	7.58	9.20
India + All states	5.72	7.29	10.56
India + All states + Google trend	5.30	7.15	8.62

and early May of 2021. The strongest Coronavirus wave to date hit India during these weeks, focusing all attention and resources on itself and away from virtually everything else. Therefore, during that time, Tuberculosis was most likely either overlooked or misdiagnosed as Coronavirus infection, leading to a sharp decline in cases recorded for TB cases each week. Figure 5 presents a comparison between the Coronavirus and tuberculosis trends during the year 2021. During the 12th week of the year (late April), COVID-19 cases in India increased at a fast rate, which led to a drop in Tuberculosis cases due to both an actual drop in cases as a result of the



(a) India +All States + Google Trend 1 week ahead

(b) India +All States + Google Trend 4 weeks ahead

(c) India +All States + Google Trend 8 weeks ahead

Fig. 4: Forecasting using India National Data + All States' data + Google Search Trends as additional Parameters

lock-downs as well as the lack of testing/reporting in this time period.

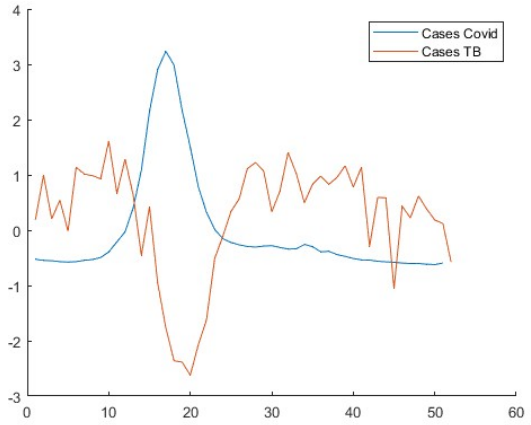


Fig. 5: Covid vs TB

Once that Coronavirus wave came to an end, the regular trend for TB reporting was restored as the graphs show, but the predictions after that valley (except in figure 4 - more on that soon) tend to be less accurate. Since the outlier is affecting the results tremendously, using a technique to treat this outlier that does not require getting rid of relevant data was able to enhance the model.

B. Outlier treatment and correction

The technique employed consists of predicting the outlier zone using the previous section as the training data set, and hence once that outlier portion is corrected, running the model again for the whole data set assuming the correction was the original data. This technique is effective since instead of getting rid of data points that might come in handy for the training of the model, it replaces the gap in the data with values that make sense in the context of the data by forecasting based on the previous behaviour the data presented. The details of this technique are the subject of another paper by the authors still under publication at the time of this writing.

It is important to note that this technique was only applied to the data set that presented the best results in table I and depicted in figure 4 - using national data as well as data from all states individually and google trends as parameters. As can be seen in table II the model does better (as noted by it's lower MAPE) at following the behaviour of the reported data, now that the outlier zone is not as present as it was before.

TABLE II: Corrected Results

data set Entry / # of weeks ahead	1	4	8
India + All states + Google trend	5.01	6.21	7.88

V. Future Steps

A. Expanding Beyond India

Many other nations struggle with Tuberculosis other than India, such as China, Nigeria, South Africa, Bangladesh, and others [2]. Having more countries added to the investigation is not only beneficial from the point of view of adding more data points to examine how TB spreads through a population, but also it can provide global context. Pandemics and epidemics are not limited by borders in the modern age, and an outbreak in one country can lead to spread in neighboring and far off nations as is still in recent memory with the spread of COVID-19. This would strengthen the data sets and potentially even improve the accuracy of the prediction by the data becoming more granular and having less gaps.

B. Comparison to Alternatives to LSTMs

Although LSTMs were selected as a basis for the model after its strong performance on predicting influenza and COVID-19 cases, direct comparison of the model's accuracy to alternatives proposed by other papers still remains. A direct comparison of various models using the data gathered in this paper would provide insights into accurately and efficiently producing forecasts, as well as examining how well models generalize to other cases outside of the ones they were trained

upon. For example, despite the ARIMA model achieving good results in China, the model might not generalize to a different domain such as cases in India.

C. Effects of Overlapping Pandemics

Although initial investigation was performed into the effects of the COVID pandemic on the spread of TB, there is further research to be done on the interaction of overlapping pandemics. Specifically, cases where individuals were infected with multiple diseases could produce sever volatility in relatively low-incidence diseases such as TB with patients having a higher mortality risk. Comparing the spread of the recent monkeypox outbreak could for instance provide further insights that are not focused on COVID, providing a more general view of the interaction between TB and other diseases.

D. Studying if Social Media Data can be useful for the model

As was done with eVision's influenza forecaster [19] social media mentions of TB can be scraped from Twitter and Reddit and added to the LSTM as additional features. It is likely that this will only lead to negligible improvements similar to eVision's influenza forecaster model, however the nonseasonal nature of TB outbreaks in comparison to influenza may fair better from social media mentions.

VI. Conclusion

Tuberculosis is not yet a disease of the past, and so as long as it continues to plague developing nations there will be a need for fast diagnosis and analysis of new outbreaks. This paper examined the use of machine learning techniques, adapting a previously successful influenza and COVID-19 forecast model to TB in India and examined what feature set would produce the best results. The eVision model clearly demonstrated the ability to forecast in an entirely new domain with the support of historical TB data, case numbers within different states, and data extracted from google trends on tuberculosis related keywords. The effect of the mid-2020 COVID-19 surge in India on the tuberculosis case numbers, namely, the depressed rates, was examined and prompted the further use of an LSTM model to make up for the disruption with likely COVID-absent behavior. More research is required in order to elevate the tuberculosis model, as it is currently unable to match the accurate 8 week in advance prediction that can be made on influenza and COVID-19 in the United States. Regardless, the forecast provided now can yield helpful warnings about the spread of TB in India which can facilitate healthcare professionals' plans and reactions to a new outbreak. And lastly, the addition of more countries' outbreak data will further enhance the predictive ability of the model beyond India.

Acknowledgment

Many thanks are due to Ben Dority, the New-Technology Senior Director at Cepheid Inc. for inspiring the project and supporting it throughout development, as well as to Eva Bouzos, SCU Masters Program Alum and Oncology R&D Research Associate at Cepheid Inc. for her continued support of the project and liaising between the team and Cepheid scientists. Also to Prashanth Asuri, director of SCU's Healthcare Innovation and Design Program for obtaining financial support from Cepheid Inc for the project. And to Santa Clara University's Frugal Innovation Hub as well as the Departments of Bioengineering, Computer Science & Engineering, Information Science & Analytics, and Mathematics & Computer Science for their continued support of the project.

References

[1] P. Sudre, G. Ten Dam, and A. Kochi, "Tuberculosis: a global overview of the situation today." *Bulletin of the World Health Organization*, vol. 70, no. 2, p. 149, 1992.

[2] W. H. Organization. (2021) Tuberculosis. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[3] ——. (2022) Tb statistics india. [Online]. Available: https://tbfacts.org/tb-statistics-india

[4] ——. (2021) Tuberculosis in china. [Online]. Available: https://www.who.int/china/health-topics/tuberculosis

[5] K. Thiagarajan, "Why is india having a covid-19 surge?" 2021.

[6] C. T. Sreeramareddy, Z. Z. Qin, S. Satyanarayana, R. Subbaraman, and M. Pai, "Delays in diagnosis and treatment of pulmonary tuberculosis in india: a systematic review," *The International Journal of Tuberculosis and Lung Disease*, vol. 18, no. 3, pp. 255–266, 2014.

[7] V. Kumar, A. Singh, M. Adhikary, S. Daral, A. Khokhar, and S. Singh, "Seasonality of tuberculosis in delhi, india: a time series analysis," *Tuberculosis research and treatment*, vol. 2014, 2014.

[8] K. Wang, C. Deng, J. Li, Y. Zhang, X. Li, and M. Wu, "Hybrid methodology for tuberculosis incidence time-series forecasting based on arima and a nar neural network," *Epidemiology & Infection*, vol. 145, no. 6, pp. 1118–1129, 2017.

[9] A. Azeez, D. Obaromi, A. Odeyemi, J. Ndege, and R. Muntabayi, "Seasonality and trend forecasting of tuberculosis prevalence data in eastern cape, south africa, using a hybrid model," *International journal of environmental research and public health*, vol. 13, no. 8, p. 757, 2016.

[10] R. Chowdhury, A. Mukherjee, S. Naska, M. Adhikary, S. K. Lahiri *et al.*, "Seasonality of tuberculosis in rural west bengal: A time series analysis," *International Journal of Health & Allied Sciences*, vol. 2, no. 2, p. 95, 2013.

[11] N. Shaghaghi, A. Calle, and G. Kouretas, "Influenza forecasting," in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 339–341.

[12] ——, "Expanding evision's scope of influenza forecasting," in *2020 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE, 2020, pp. 1–10.

[13] N. Shaghaghi, A. Calle, G. Kouretas, S. Karishetti, and T. Wagh, "Expanding evision's granularity of influenza forecasting," in *Wireless Mobile Communication and Healthcare: 9th EAI International Conference, MobiHealth 2020, Virtual Event, November 19, 2020, Proceedings*. Springer Nature, p. 227.

[14] N. Shaghaghi, A. Calle, G. Kouretas, J. Mirchandani, and M. Castillo, "evision: Epidemic forecasting on covid-19," *Current Directions in Biomedical Engineering*, vol. 7, no. 2, pp. 839–842, 2021.

[15] N. Shaghaghi, S. Karishetti, and N. Ma, "Interplay of influenza a/b subtypes and covid-19," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*. IEEE, 2021, pp. 1–5.

[16] N. Shaghaghi, A. Calle, and Y. Qian, "evision: Influenza forecasting using cdc, who, and google trends data," in *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*. IEEE, 2020, pp. 38–45.

[17] C. Olah. (2015) Understanding lstm networks. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs

[18] S. H. To. (2018) Z-score: Definition, formula and calculation. [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/z-score

[19] N. Shaghaghi, Y. Kamdar, R. Huang, A. Calle, J. Mirchandani, and M. Castillo, "Attempts at enhancing evision's influenza forecasting using social media," in *2022 14th Biomedical Engineering International Conference (BMEiCON)*. IEEE, 2022, pp. 1–5.