Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 12 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i12.56193

PAPER

Artificial Intelligence-Based Surveillance of Tuberculosis in South Africa Using Google Trends Data

Nqobile S. Hlatshwayo, Seun O. Olukanmi (✉)

University of the Witwatersrand, Johannesburg, South Africa

seun.olukanmi@wits.ac.za

ABSTRACT

Tuberculosis (TB) remains a significant global public health challenge and the deadliest infectious disease, according to the World Health Organization (WHO) Global TB Report of 2024. This study explores integrating Google Trends (GT) data with machine learning (ML) methods to forecast TB incidence in South Africa. Pearson correlation analysis identified eight TB-related search terms with moderate to strong correlations to official surveillance data from the National Institute for Communicable Diseases (NICD) between 2012–2021. Four ML models were compared using rolling-window cross-validation: partial least squares (PLS), LASSO regression, support vector machine (SVM), and long short-term memory (LSTM) networks. The PLS model achieved superior performance, significantly outperforming more complex deep learning approaches. These findings demonstrate that simpler linear models can effectively leverage GT data to complement traditional TB surveillance systems in South Africa.

KEYWORDS

tuberculosis (TB), disease surveillance, machine learning, Google Trends (GT), infoveillance, digital epidemiology, artificial intelligence, disease surveillance, epidemiology, South Africa, public health

1 INTRODUCTION

1.1 Evolution of tuberculosis in South Africa

The history of tuberculosis (TB) in South Africa is marked by significant trends and key developments. According to Karim *et al.* [1], TB was a relatively rare disease before the 20th century, primarily affecting European settlers and individuals in close contact with them. However, TB prevalence surged dramatically in the late 19th and early 20th centuries, driven by factors such as urban overcrowding, labour migration, fragmented healthcare systems under apartheid, the emergence of the HIV epidemic, and delayed government responses. It was only in 2004 that TB was declared a national emergency.

Hlatshwayo, N. S., Olukanmi, S. O. (2025). Artificial Intelligence-Based Surveillance of Tuberculosis in South Africa Using Google Trends Data. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(12), pp. 94–105. https://doi.org/10.3991/ijoe.v21i12.56193

Article submitted 2025-04-23. Revision uploaded 2025-07-10. Final acceptance 2025-07-10.

In the post-World War II era, the South African government introduced various TB control measures, including the rifampicin-based Directly Observed Therapy, short course (DOTS) strategy recommended by the WHO [1]. These measures initially reduced TB notification rates during the 1980s. However, the rapid escalation of the HIV epidemic in the late 1990s reversed these gains, leading to a sharp rise in TB cases.

Currently, South Africa continues to face one of the world's highest TB burdens, with an estimated incidence rate of 427 cases per 100,000 population in the past year, as reported by the *WHO Global TB Report 2024* [2].

1.2 Traditional TB surveillance methods in South Africa

Despite considerable progress in expanding TB diagnostic and treatment services, South Africa's TB surveillance system still encounters notable challenges, including incomplete case notifications, data quality deficiencies, and the absence of real-time monitoring [3]. Traditional TB surveillance methods often experience reporting delays of weeks or even months, which impede timely outbreak detection and response [4]–[5].

May, Chretien, and Pavlin [6] emphasise that conventional surveillance systems are resource-intensive, necessitating considerable investments in personnel, equipment, and laboratory space. This is especially problematic in low-income regions or during financial constraints, making it particularly relevant in the South African context [6].

These shortcomings hinder efforts to accurately measure the true scale of the TB epidemic, track transmission patterns, and evaluate the real-world impact of interventions. Furthermore, they obstruct South Africa's progress towards its national goal of eradicating TB by 2035 [7]. Consequently, these limitations highlight the importance of leveraging innovative data sources and artificial intelligence (AI) techniques to enhance infectious disease surveillance, including TB.

This study addresses three questions: (1) Which TB-related Google search terms exhibit stable correlations with incidence rates in South Africa? (2) How do various well-known models (linear and non-linear) perform in predicting TB trends? (3) Can Google Trends (GT) data bridge surveillance gaps in resource-limited settings?

2 RELATED WORK

Information gathering for public health through Internet-Based Sources (IBS), referred to as "infodemiology" or "infoveillance", has gained traction since the early 20th century [5]. Early studies by Polgreen *et al.* [8] and Ginsberg *et al.* [9] demonstrated that search data could predict influenza-like illnesses, inspiring further research, including work by Brownstein *et al.* [10], Corley and Mikler [11], Valdivia and Monge [12], Pelat *et al.* [13], and Hulth *et al.* [14].

Amusa *et al.* highlighted the use of GT data, often combined with other digital sources, to track and predict diseases such as influenza [8], [15], [16] and COVID-19 [5], [17]. These studies reveal that online search patterns can act as early indicators of health trends and disease spread.

Zhou *et al.* [5] pioneered the use of GT data for TB surveillance in the U.S., employing a non-stationary dynamic system model to capture disease transmission

dynamics and seasonal patterns. Their model provided TB estimates up to 12 weeks ahead of CDC reports, showcasing its potential for early outbreak detection.

Building on this, Santillana *et al.* [16] applied a machine learning (ML) ensemble using GT data, Twitter posts, hospital records, and participatory surveillance systems to now-cast and forecast influenza activity, outperforming individual data sources.

Deep learning approaches have also shown promise. Prasanth *et al.* [18] developed a hybrid model combining Grey Wolf Optimisation and LSTM networks to forecast COVID-19 spread using GT data, outperforming all other models in the study. Conversely, Ayyoubzadeh *et al.* [17] found that simpler models, like linear regression, performed better in predicting daily COVID-19 cases in Iran.

In TB surveillance particularly, Santos [19] used GT data in Portugal to test various statistical and ML models, finding that partial least squares (PLS) regression outperformed more complex models, achieving a correlation of 0.7 between predicted and actual TB cases. This aligns with Pavlin's observation [20] that infoveillance can effectively complement traditional methods using straightforward tools and minimal resources.

While previous studies highlight the potential of AI-based disease surveillance, they also emphasise the need for research in high-burden regions and diseases like TB. As suggested in [6], infoveillance systems should enhance existing public health infrastructure while leveraging prior regional research.

Recent African studies have demonstrated this potential: Nsoesie *et al.* [29] successfully applied GT with ML to forecast influenza-like illness in Cameroon (R^2 = 0.78 – 0.88), while Bragazzi and Mahroum [30] effectively monitored Madagascar's 2017 plague outbreak using search data, both highlighting the feasibility of digital surveillance in resource-limited African settings.

To date, no studies have examined the use of GT data with AI-based methods for TB surveillance in South Africa. Addressing this gap while building on existing work could provide novel insights and significantly enhance TB surveillance and control in the region.

3 METHODOLOGY

3.1 Data and data preprocessing

This study received a waivered ethics clearance by the Human Research Ethics Committee of the University of Witwatersrand, Johannesburg. The data used in the study covered a period of 40 quarters, comprising quarterly data from the first quarter of 2012 to the fourth quarter of 2021. There were two primary data sources:

- **Official TB incidence data:** The official TB quarterly incidence data in South Africa were retrieved from the National Institute for Communicable Diseases (NICD) online TB dashboard, which is the last known public official data. The NICD operates as a division of the National Health Laboratory Service and serves as a resource for knowledge and expertise in infectious diseases to the South African government, healthcare providers, and the public [21].
- **Google Trends data:** GT data were collected to complement official statistics, leveraging South Africa's high Google usage rate (95% of internet users) [3]. We identified and tracked 27 TB-related search terms across five categories: Name, Symptom, Cause, Diagnosis and Related Disease. GT allows you to refine

the results based on period, category, type of search and geolocation. For the purposes of this study, search volume data was extracted for the previously specified period of 2012–2021, with the location set to South Africa. The type of search is set to 'Web search', and the categories explored were 'All Categories' as well as 'Health'. This allowed us to retrieve weekly TB search data in South Africa for the research period.

Data preprocessing involved several key steps: handling missing values using K-nearest neighbours (KNN) imputation, chosen for its ability to preserve relationships between features by considering multiple neighbours and its suitability for time series data where values tend to be similar to neighbouring time points [22]; normalising TB incidence rates to the [0, 1] range while maintaining GT data in its original 0–100 scale; and aggregating weekly GT data to quarterly intervals for temporal alignment. The dataset was split into training (80%) and testing (20%) sets as recommended in [23]. This comprehensive preprocessing approach ensured the data was optimally prepared for model development while preserving the inherent characteristics of both TB incidence and GT data.

3.2 Feature selection and correlation analysis

Working from our initial list of 27 search terms, we identified the most relevant GT search terms for TB surveillance through a comprehensive correlation analysis. This involved calculating Pearson correlation coefficients between each search term (and all its variants in terms of subcategories) with TB incidence rates. The analysis was performed separately for each year from 2012–2021 to assess temporal stability. For each correlation, we report the corresponding *p*-value and 95% confidence interval (CI), derived from a two-tailed test with significance level $\alpha = 0.05$.

Feature selection was based on overall correlations across the full study period ($n = 40$ quarters) to ensure adequate statistical power, supplemented by temporal consistency analysis across individual years. While yearly correlations provided insights into temporal stability, the limited sample size per year ($n = 4$) necessitated focusing on overall relationships for statistical inference. Terms needed to demonstrate a moderate to strong positive correlation with TB incidence rates, defined as $r \geq 0.4$.

3.3 Model development and evaluation

The time-dependent nature of time-series data requires careful consideration during model validation to preserve temporal relationships. Conventional cross-validation methods are unsuitable in this scenario because they can breach temporal dependencies by using future data to predict past events [24]. To address this issue, a rolling window strategy with time-series cross-validation was employed [25]. This approach ensures the temporal integrity of the data while facilitating robust model validation.

Our implementation utilised a rolling window methodology where the models were initially trained on twelve quarters of historical data and tested on the subsequent quarter. This window then "rolled" forward one quarter at a time, maintaining the twelve-quarter training period while generating predictions for

each new quarter. This approach was particularly relevant for our study period (2012–2021), which included significant events such as the COVID-19 pandemic, allowing us to assess model performance across both stable and volatile periods.

We implemented and compared four distinct models:

1. Least absolute shrinkage and selection operator (LASSO) regression
   - The model uses the Lasso CV class from the sklearn.linear_model library in Python.
   - The class automatically performs hyperparameter tuning using 5-fold cross-validation to determine the best value for the regularisation parameter α.
2. Partial least squares (PLS) regression
   - The model was implemented using the PLSRegression class from the sklearn.cross_decomposition library in Python.
   - Hyperparameter tuning revealed an optimal n_component parameter of 2.
3. Support vector machine (SVM)
   - Implemented using scikit-learn's SVR class.
   - We used GridSearchCV from sklearn.model_selection for hyperparameter tuning.
4. Long short-term memory (LSTM)
   - Implemented in Keras, running on top of a TensorFlow backend in Python
   - Consists of two LSTM layers (with 64 and 32 units, respectively), each followed by a dropout layer to prevent overfitting
   - The dropout rate and learning rate for the Adam optimiser were set to 0.2 and 0.01, respectively (determined through hyperparameter tuning)

Performance was assessed using multiple metrics: Mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2). To assess the statistical significance of performance differences between models, paired t-tests were conducted comparing the MSE of each model pair.

4 RESULTS AND DISCUSSION

4.1 Correlation analysis

The correlation analysis revealed significant relationships between GT search patterns and TB incidence rates. After considering all permutations and concatenations of search terms and subcategories, we had a total of 36 search terms. Out of these 36 queries, 11 of them did not have data. For most of these missing values we were able to use KNN imputation as suggested in [22], but in cases where this was infeasible, we discarded the respective query. This process left us with 27 search terms, which we then used in the correlation analysis. The overall correlation analysis revealed eight search terms that had a moderate to strong significant positive overall correlation with the TB incidence rates. Figure 1 presents these overall correlations with 95% confidence intervals, while detailed yearly correlation patterns are provided in Table 1.

Comorbidity-related searches emerged as robust predictors, with diabetes-related terms demonstrating the highest overall correlations. This pattern likely emerges from increasing diabetes prevalence in urban populations, heightened clinical awareness of TB-diabetes comorbidity, and patient education initiatives highlighting diabetes as a TB risk factor.

Disease-specific searches for tuberculosis terms demonstrated moderate to strong correlations. This positive correlation could indicate that people experiencing TB symptoms or those concerned about TB actively seek information online before or after seeking medical attention.

HIV-related searches also exhibit moderate to strong correlations, which is particularly significant in the South African context as it directly reflects the established HIV-TB coinfection pattern in South Africa. This suggests that people who are HIV-positive or concerned about HIV are also actively seeking information about TB, indicating awareness of the connection between these conditions.

Overall, we find that the correlation analysis results are analogous to those published by Fudholi and Fikri [26] in Indonesia as well as Santos [19] in Portugal. There is a continuous pattern of high positive correlations between the TB incidence in each nation and the GT trend statistics.

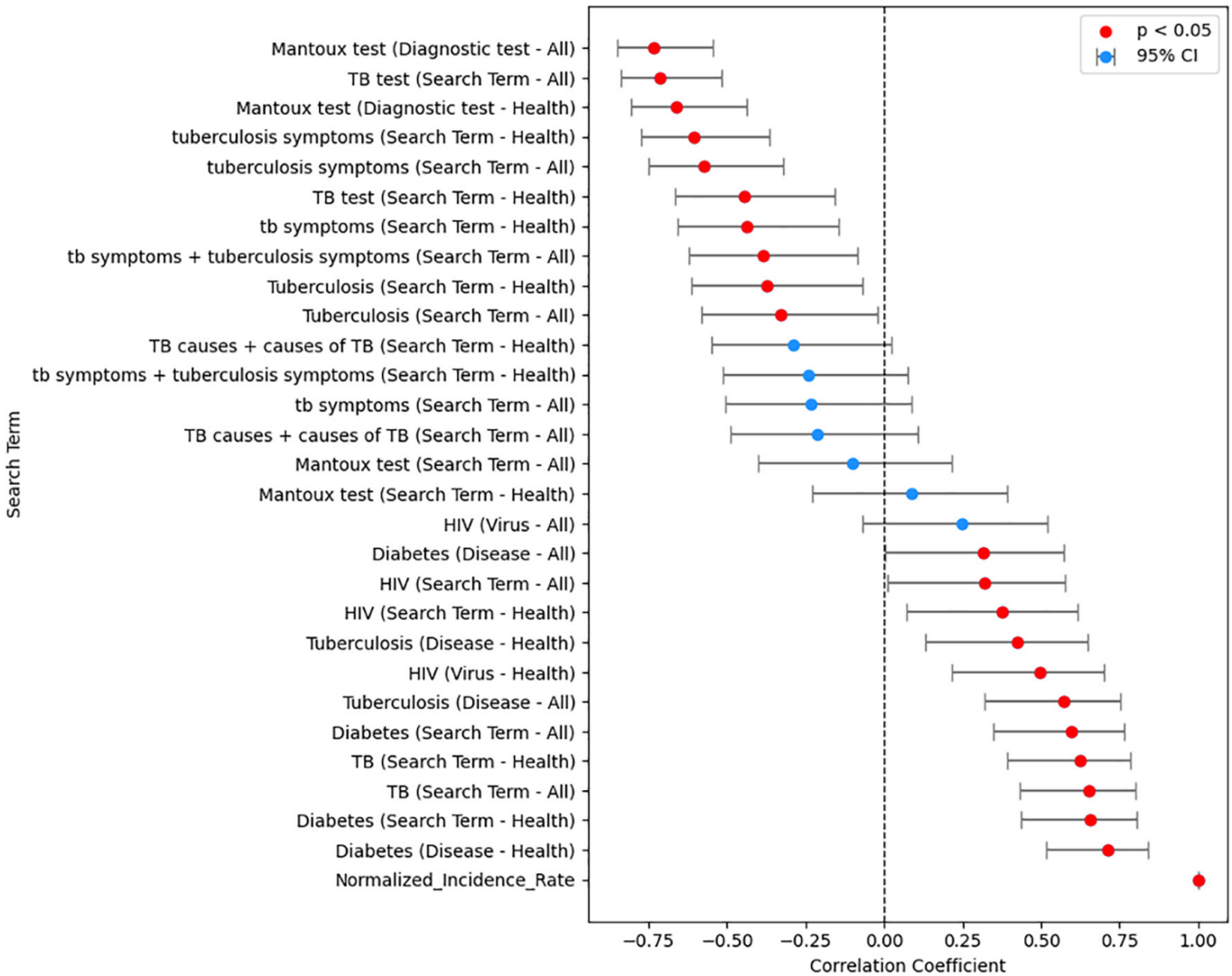

Fig. 1. Overall Pearson correlations between GT search term and TB incidence

Note: Each point represents the correlation coefficient (r) between a search term and TB incidence over the entire study period. The horizontal error bars indicate the 95% confidence interval. Red points denote statistically significant correlations ($p < 0.05$); the vertical dashed line indicates $r = 0$.

Table 1. Yearly Pearson correlations of search terms with TB incidence rate

No	Search Term	Category	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
1	Diabetes (Disease)	All	0.22	–0.01	0.83	0.27	–0.03	0.6	0.59	0.59	**–0.98**	0.27
2	Diabetes (Disease)	Health	0.17	0.13	**0.98**	0.37	0.07	0.61	0.73	0.53	**–0.96**	0.21
3	Diabetes (Term)	All	0.18	0.20	0.76	0.47	0.73	0.64	0.51	0.44	**–1.0**	–0.032
4	Diabetes (Term)	Health	0.2	0.21	0.75	0.6	0.88	0.87	0.5	0.62	**–0.96**	–0.28
5	HIV (Term)	All	0.10	0.12	0.76	0.77	**0.95**	0.30	0.68	0.84	0.87	–0.19
6	HIV (Term)	Health	0.06	0.30	0.8	0.64	**0.98**	0.34	0.92	0	0.87	0.042
7	HIV (Virus)	All	0.01	0.0	0.76	0.77	0.93	0.28	0.33	0.63	0.91	–0.13
8	HIV (Virus)	Health	0.02	0.30	0.76	0.71	**0.98**	0.31	0.22	0.72	0.88	0.062
9	Mantoux test (Diagnosis)	All	0.21	0.78	0.083	–0.11	–0.1	0.23	0.4	0.052	0.54	–0.42
10	Mantoux test (Diagnosis)	Health	–0.8*	–0.8	**0.97**	–0.025	–0.44	0.49	–0.67	–0.55	–0.04	–0.33
11	Mantoux test (Diagnosis)	All	0.43*	0.43*	0.43	0.74	–0.103	0.232	–0.16	–0.36	0.16*	0.16
12	TB (Term)	All	–0.15	–0.2	0.74	0.22	0.38	–0.24	0.23	0.36	0.51	–0.46
13	TB (Term)	Health	0.23	–0.31	0.79	0.64	0.81	0.25	0.56	0.71	0.49	–0.54
14	TB causes + causes of TB (Term)	All	–0.27	–0.37	0.78	0.93	0.86	0.17	0.65	–0.019	0.26	0.1
15	TB causes + causes of TB (Term)	Health	–0.79	0.73	0.75	–0.025	0.41	0.42	–0.46	0.41	–0.54	0.51
16	TB symptoms + tuberculosis symptoms (Term)	All	0.16	0.37	0.78	0.51	0.93	**0.95**	**0.96**	0.78	0.92	0.08
17	TB symptoms + tuberculosis symptoms (Term)	Health	0.31	0.91	0.81	0.58	0.88	0.86	0.72	0.78	0.92	–0.74
18	TB test (Term)	All	0.28*	0.28*	0.28	0.61	0.81	0.13	0.80	–0.15	0.88	**0.96**
19	TB test (Term)	Health	0.28*	0.28*	0.28	0.61	0.81	0.66	0.78	0.82	**0.98**	0.35
20	TB symptoms (Term)	All	0.23	0.62	0.69	0.3	0.92	0.81	0.86	0.54	0.87	0.11
21	TB symptoms (Term)	Health	–0.26	0.42	0.46	0.18	0.82	0.94	0.94	0.48	0.87	0.23
22	Tuberculosis (Disease)	All	0.21	–0.27	0.86	0.53	0.83	0.017	0.86	0.75	0.052	–0.41
23	Tuberculosis (Disease)	Health	0.25	–0.3	0.89	0.5	0.81	0.2	0.87	0.74	0.12	0.055
24	Tuberculosis (Term)	All	0.37	–0.4	0.88	0.22	0.75	–0.19	0.29	0.77	–0.16	–0.32
25	Tuberculosis (Term)	Health	0.25	–0.49	**0.996**	0.46	0.75	–0.10	0.37	0.63	–0.099	–0.53
26	Tuberculosis symptoms (Term)	All	0.33	0.20	0.099	0.21	0.31	–0.34	0.76	0.52	–0.68	–0.30
27	Tuberculosis symptoms (Term)	Health	0.56	0.56*	0.63	0.81	–0.5	–0.73	0.73	–0.73	–0.57	–0.74

Note: Correlation coefficients (r) between GT search terms and TB incidence rates for each year from 2012–2021. The bold font numbers indicate statistically significant values ($p < 0.05$), and * indicates KNN imputation.

4.2 Model performance analysis

The performance comparison of the four ML models implemented in this study reveals a clear hierarchy in their predictive capabilities for TB surveillance in South Africa. Figure 2 presents the visualisation of actual versus predicted TB incidence rates across all four models, revealing distinct patterns in their predictive capabilities.

The results of our study demonstrate strong predictive accuracy and align with existing literature. The PLS model outperformed other approaches, significantly surpassing the performance of the LSTM. This outcome is consistent with Santos [19], whose research on TB surveillance in Portugal also identified PLS as the most effective model for GT-based predictions, though absolute metrics differed due to variations in data scaling. This superior performance likely reflects PLS's ability to handle collinear predictors and extract meaningful latent variables from correlated GT features, making it particularly well-suited for this application where multiple search terms may reflect similar underlying health-seeking behaviours.

LASSO regression performed nearly as well, showing solid reliability in identifying medium-term trends, although its precision declined slightly at key epidemiological shifts. The SVM yielded moderate results, often overestimating cases, which reduced its reliability.

The LSTM network's underperformance warrants careful consideration. Several factors likely contributed to this outcome: (1) the relatively limited training data (40 quarters) and low-dimensional features. Deep learning typically requires larger datasets to capture complex patterns—a constraint noted in Ayyoubzadeh *et al.* [17] for COVID-19 predictions. (2) TB's endemic rather than epidemic dynamics may not require the sophisticated temporal modelling that LSTMs provide; and (3) potential overfitting due to the model's complexity relative to available training data. Additionally, the stable nature of TB transmission patterns in South Africa may be better captured by simpler approaches that avoid the noise introduction common in over-parameterised models.

The statistical significance testing revealed that PLS significantly outperformed both SVM ($t = -2.478$, $p = 0.020$) and LSTM ($t = -2.686$, $p = 0.012$). LASSO also demonstrated significant superiority over LSTM ($t = -2.124$, $p = 0.043$). Notably, no significant difference was found between PLS and LASSO ($t = -1.654$, $p = 0.110$), indicating that both approaches provide competitive performance. The comparison between LASSO and SVM approached significance ($t = -1.843$, $p = 0.077$), while SVM and LSTM showed no significant difference ($t = -1.455$, $p = 0.158$). These results statistically validate the superior performance of simpler linear models over complex deep learning approaches for TB surveillance using GT data.

4.3 Limitations of the study

A key challenge in using GT for disease surveillance is the digital divide, which refers to the gap between those who have access to and use the Internet and those who do not [27]. However, South Africa's growing digital connectivity, with a 21% increase in the Internet penetration in just five years [28], indicates substantial potential for leveraging online data for public health monitoring. Similarly, the social media user base is steadily expanding, creating promising opportunities for social media-based disease surveillance. As digital adoption continues to grow

across all segments of the population, the utility and representativeness of these data sources will continue to improve.

The diversity of languages used online in South Africa is a unique challenge but also an opportunity to develop more inclusive and representative data collection and analysis approaches. By accounting for the country's 12 official languages, researchers can capture a broader range of online health-seeking behaviour and enhance the accuracy and generalisability of their disease monitoring models.

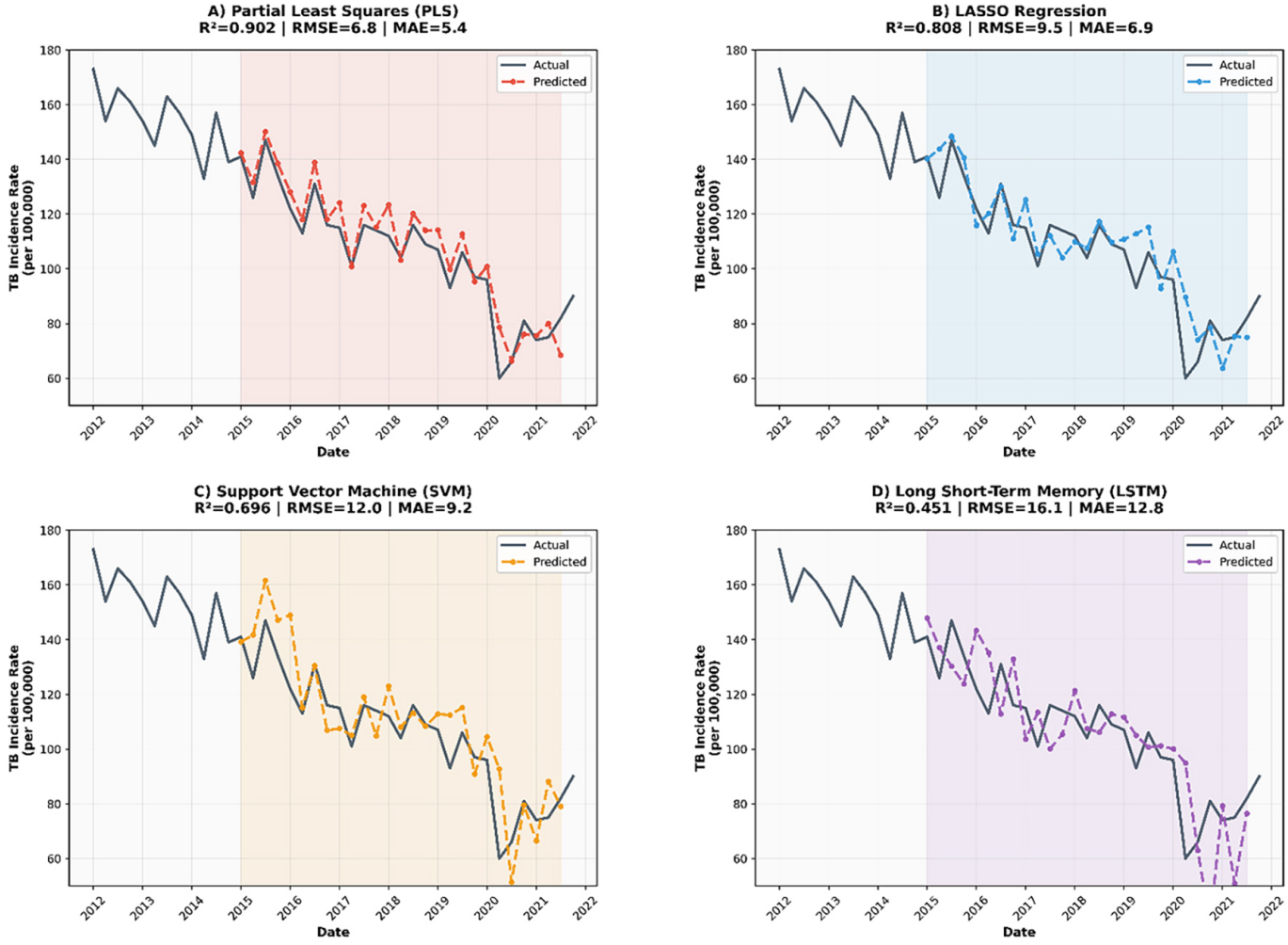

Fig. 2. Comparison of ML model performance for TB incidence prediction

Note: Four different models—PLS, LASSO, SVM and LSTM—are evaluated for predicting TB incidence rates over time. Black lines represent actual TB incidence rates (per 100,000 population), while coloured lines show model predictions.

5 CONCLUSION

Tuberculosis remains a significant public health challenge in South Africa, requiring innovative approaches to improve surveillance and control. This study highlights the potential of GT data combined with ML models to enhance TB monitoring.

The analysis revealed strong correlations between Google search patterns—particularly symptom-related and HIV-associated queries—and TB incidence rates,

suggesting their potential as early indicators. Among the four ML models evaluated, PLS regression emerged as the best performer with an R^2 of 0.902 and superior error metrics, outperforming complex models like LSTM networks. This demonstrates that simpler models can effectively capture the relationship between GT data and TB incidence, offering computational efficiency and interpretability.

Future advancements should focus on multilingual natural language processing to analyse diverse linguistic data and explore ensemble approaches that integrate multiple data sources for comprehensive surveillance. While GT cannot replace traditional TB monitoring systems, it provides a cost-effective, scalable, and complementary tool for early warning and real-time monitoring.

These findings support South Africa's Strategic Plan to end TB by 2035 and underline the role of digital tools in accelerating progress towards this national and global health goal.

6 REFERENCES

[1] S. S. A. Karim, G. J. Churchyard, Q. A. Karim, and S. D. Lawn, "HIV infection and tuberculosis in South Africa: An urgent need to escalate the public health response," *The Lancet*, vol. 374, no. 9693, pp. 921–933, 2009. https://doi.org/10.1016/S0140-6736(09)60916-8

[2] World Health Organization. *Global Tuberculosis Report 2024*. License: CCBY-NC-SA3.0 IGO. 2024.

[3] G. J. Churchyard *et al.*, "Tuberculosis control in South Africa: Successes, challenges and recommendations: Tuberculosis control-progress towards the millennium development goals," *South African Medical Journal*, vol. 104, no. 2, pp. 244–248, 2014. https://doi.org/10.7196/SAMJ.7689

[4] C. T. Sreeramareddy, K. V. Panduru, J. Menten, and J. Van Den Ende, "Time delays in diagnosis of pulmonary tuberculosis: A systematic review of literature," *BMC Infectious Diseases*, vol. 9, pp. 1–10, 2009. https://doi.org/10.1186/1471-2334-9-91

[5] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by an analyzing Google Trends," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 8, pp. 2247–2254, 2011. https://doi.org/10.1109/TBME.2011.2132132

[6] L. May, J.-P. Chretien, and J. A. Pavlin, "Beyond traditional surveil lance: Applying syndromic surveillance to developing settings – opportunities and challenges," *BMC Public Health*, vol. 9, pp. 1–11, 2009. https://doi.org/10.1186/1471-2458-9-242

[7] South African National Department of Health. *Strategic Plan 2020/21 – 2024/25*. Pretoria, South Africa, 2020.

[8] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson, "Using internet searches for influenza surveillance," *Clinical Infectious Diseases*, vol. 47, no. 11, pp. 1443–1448, 2008. https://doi.org/10.1086/593098

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009. https://doi.org/10.1038/nature07634

[10] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection—harnessing the web for public health surveillance," *The New England Journal of Medicine*, vol. 360, no. 21, p. 2153, 2009. https://doi.org/10.1056/NEJMp0900702

[11] C. D. Corley and A. R. Mikler, "A computational framework to study public health epidemiology," in *2009 International Joint Conference on Bioinformatics, Systems Biology, and Intelligent Computing*, 2009, pp. 360–363. https://doi.org/10.1109/IJCBS.2009.83

[12] A. Valdivia and S. Monge-Corella, "Diseases tracked by using Google Trends, Spain," *Emerging Infectious Diseases*, vol. 16, no. 1, p. 168, 2010. https://doi.org/10.3201/eid1601.091308

[13] C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, and A.-J. Valleron, "More diseases tracked by using Google Trends," *Emerging Infectious Diseases*, vol. 15, no. 8, p. 1327, 2009. https://doi.org/10.3201/eid1508.090299

[14] A. Hulth, G. Rydevik, and A. Linde, "Web queries as a source for syndromic surveillance," *PLoS ONE*, vol. 4, no. 2, p. e4378, 2009. https://doi.org/10.1371/journal.pone.0004378

[15] F. S. Lu *et al.*, "Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the Boston Metropolis," *JMIR Public Health and Surveillance*, vol. 4, no. 2, p. e8950, 2018. https://doi.org/10.2196/publichealth.8950

[16] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PloS Computational Biology*, vol. 11, no. 4, p. e1004513, 2015. https://doi.org/10.1371/journal.pcbi.1004513

[17] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study," *JMIR Public Health and Surveillance*, vol. 6, no. 6, p. e18828, 2020. https://doi.org/10.2196/18828

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

[19] L. G. Santos, "Surveillance of tuberculosis by analysing Google Trends," M. S. Thesis, Universidade Católica Portuguesa, Lisbon, Portugal, 2023.

[20] J. A. Pavlin, "Asia: Syndromic surveillance in Asia–opportunities and challenges," *International Society for Disease Surveillance. Global Outreach Committee Newsletter*, 2008.

[21] South African national health laboratory service. *NHLS Strategic Plan 2020/21 – 2024/25*. Johannesburg, South Africa, 2020.

[22] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001. https://doi.org/10.1093/bioinformatics/17.6.520

[23] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts, 2018.

[24] W.-J. Chen, J.-J. Yao, and Y.-H. Shao, "Volatility forecasting using deep neural network with time-series feature embedding," *Economic Research-Ekonomska Istraživanja*, vol. 36, no. 1, pp. 1377–1401, 2023. https://doi.org/10.1080/1331677X.2022.2089192

[25] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012. https://doi.org/10.1016/j.ins.2011.12.028

[26] D. H. Fudholi and K. Fikri, "Towards an effective tuberculosis surveillance in Indonesia through Google Trends," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 5, no. 4, pp. 299–308, 2020. https://doi.org/10.22219/kinetik.v5i4.1114

[27] Z. Langa, P. Conradie, and B. Roberts, "Slipping through the net: Digital and other communication divides within South Africa," in *South African Social Attitudes: Changing Times, Diverse Voices*, Cape Town, South Africa: HSRC Press, 2006, pp. 131–149.

[28] Data Reportal, *Digital 2024: South Africa*. 2024. https://datareportal.com/reports/digital-2024-south-africa?rq=SouthAfrica [Accessed: Apr. 27, 2024].

[29] E. O. Nsoesie, O. Oladeji, A. S. A. Abah, and M. L. Ndeffo-Mbah, "Forecasting influenza-like illness trends in Cameroon using Google Search data," *Scientific Reports*, vol. 11, no. 1, p. 6713, 2021. https://doi.org/10.1038/s41598-021-85987-9

[30] N. L. Bragazzi and N. Mahroum, "Google Trends predicts present and future plague cases during the plague outbreak in Madagascar: Infodemiological study," *JMIR Public Health and Surveillance*, vol. 5, no. 1, p. e13142, 2019. https://doi.org/10.2196/13142

7 AUTHORS

Nqobile S. Hlatshwayo is a Master's student in Computer Science at the University of the Witwatersrand, Johannesburg, South Africa. His research focuses on artificial intelligence applications in public health surveillance, particularly using machine learning, natural language processing, and digital epidemiology for infectious disease monitoring in multilingual African contexts (E-mail: nqobiletheconqueror@gmail.com).

Seun O. Olukanmi is presently a Lecturer in the School of Computer Science and Applied Mathematics at the University of the Witwatersrand, South Africa. She obtained her PhD from the University of Johannesburg. Her research focus encompasses artificial intelligence and data science for social good, cutting across several fields such as public health, precision agriculture, education, finance, cybersecurity, and human language processing (E-mail: seun.olukanmi@wits.ac.za).