Part 1: Literature Review

**Project: Forecasting Disease Outbreaks Using Integrated Mobility and Health Data**

#### 1. Introduction

The COVID-19 pandemic demonstrated a critical link between human mobility and the spread of infectious disease. As governments and health agencies seek to improve readiness for future outbreaks, the ability to forecast case spikes becomes a vital tool. This research is important because it explores the power of using publicly available data (mobility and health statistics) to create such an early warning system. A review of the existing literature is necessary to understand how other researchers have successfully modeled this relationship and to identify the most effective methods.

#### 2. Organization

This review is organized thematically, grouping papers into two main areas:

1. **Theme 1: Establishing the Link** - Papers that analyze the historical correlation between mobility and disease spread.
2. **Theme 2: Forecasting Models** - Papers that specifically use machine learning or deep learning to predict future outbreaks.

#### 3.Summary and Synthesis

**Theme 1: Establishing the Link**

* **Kraemer et al. (2020). “The effect of human mobility and control measures on the COVID-19 epidemic in China.”**
  + **Summary:** This foundational study used mobile phone data to track human movement in China. It found a strong, measurable association between the reduction in mobility (due to lockdowns) and a slowdown in the spread of COVID-19. Its key contribution was providing early, data-driven evidence that mobility restrictions were effective.
* **Pepe et al. (2021). “COVID-19 outbreak response: a first assessment of mobility changes in Italy following national lockdown.”**
  + **Summary:** Similar to Kraemer et al., this study analyzed Google Community Mobility data for Italy. It also found a significant correlation between reduced mobility and slowed disease transmission. Its methodology is highly relevant as it uses one of our proposed datasets (Google Mobility Reports).

**Theme 2: Forecasting Models**

* **Jiao et al. (2025). "Spatio-temporal epidemic forecasting using mobility data with LSTM networks and attention mechanisms."**
  + **Summary:** This recent paper proposes a novel deep learning model using LSTMs to forecast COVID-19 trends in Japan. It explicitly uses mobility data as a key input. Its key finding is that deep learning models like LSTMs can successfully model the complex, non-linear relationship between human movement and future case numbers, outperforming simpler models.
* **Ayyoubzadeh et al. (2020). "Predicting COVID-19 incidence through analysis of Google Trends and mobility data."**
  + **Summary:** This study also aimed to predict case numbers, but it used simpler machine learning models. It found that even basic regression models, when fed the right data (like mobility and search trends), could provide a valuable forecast. This demonstrates that a machine-learning approach is a valid baseline.

**Synthesis** The literature is in strong agreement. Papers like Kraemer et al. (2020) and Pepe et al. (2021) establish that a link exists between mobility and spread. Papers like Jiao et al. (2025) take the next step: they show how this link can be used for forecasting, specifically validating the use of LSTMs.

#### 4. Conclusion

* **Key Takeaways:** The key takeaway is that the link between mobility and disease spread is well-established, and using this data for forecasting is a viable and active area of research.
* **Our Contribution:** Our project will contribute by conducting a comprehensive, comparative study. While many papers apply one or two models, we will implement and evaluate a full suite of technologies—from a **Random Forest** baseline to established models like **LSTM** and **GRU**, and cutting-edge architectures like **TCN**, **Transformers**, and **Hybrid models**. By testing all of them on the same integrated health and mobility dataset, our project will provide a clear, practical evaluation of which model offers the best performance for this critical public health task.

#### 5. Citations

* Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & R Niakan Kalhori, S. (2020). Predicting COVID-19 incidence through analysis of Google Trends and mobility data. *Health Information Science and Systems*, 8(1), 1-11.
* Jiao, S., et al. (2025). Spatio-temporal epidemic forecasting using mobility data and deep learning. *Scientific Reports*.
* Kraemer, M. U., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., ... & Brownstein, J. S. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490), 493-497.
* Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C., & Tizzoni, M. (2020). COVID-19 outbreak response, a first assessment of mobility changes in Italy following national lockdown. *Scientific Data*, 7(1), 449.

Part 2: Data Research

#### 1. Introduction

The COVID-19 pandemic underscored the profound global challenge of rapidly spreading airborne diseases. The unpredictable nature of these outbreaks poses significant threats to public health systems, economic stability, and progress toward Sustainable Development Goal 3 (Good Health and Well-Being). The ability to accurately forecast spikes in disease cases is no longer a purely academic exercise; it is a critical component of national and global preparedness.

A thorough exploration of available data is necessary because our hypothesis rests on the complex, non-linear interactions between three distinct domains: human behavior (mobility), virus transmission (health statistics), and external interventions (government policy). A model that fails to account for all three (for example, by ignoring that a government lockdown was the *cause* of reduced mobility) will learn false relationships and fail as a predictive tool. This research phase is dedicated to identifying and validating the most robust and relevant datasets to build a model with a true understanding of the pandemic's dynamics.

#### 2. Organization

This data research is organized thematically, based on the three core pillars of data required for our predictive model. Each section details a specific dataset chosen to fulfill one of these roles. The three data categories are:

1. **Epidemiological Data (The Target):** This is the "what" we are trying to predict (e.g., future case numbers, positivity rates).
2. **Population Mobility Data (The Behavioral Predictor):** This is our primary behavioral feature, capturing how people's movements influence transmission.
3. **Government Policy Data (The Contextual Predictor):** This is the crucial contextual feature, explaining the *why* behind large-scale shifts in mobility and transmission.

#### 3. Data Description

Based on a thorough review, we have selected the following three datasets as the foundation for this project.

**Dataset 1: Epidemiological Data**

* **Data Source:** Our World in Data (OWID) COVID-19 Dataset
* **Data Format:** A single, comprehensive CSV file.
* **Data Size:** Large (over 500MB), containing global time-series data for all countries. We will filter this down to our specific country of interest (e.g., India).
* **Why Chosen & Relevance:** This is one of the most trusted and widely used datasets globally. It aggregates data directly from official sources (like Johns Hopkins University and national health agencies) into a single, easy-to-use file. Crucially, it includes not just new\_cases but also new\_tests, positive\_rate, hospitalizations, and vaccinations. This allows us to build a model on a more reliable metric than raw case counts, which were heavily skewed by testing availability.

**Dataset 2: Population Mobility Data**

* **Data Source:** Google Community Mobility Reports
* **Data Format:** A collection of CSV files, one for each country.
* **Data Size:** Moderate. The global file is large, but the country-specific file is manageable.
* **Why Chosen & Relevance:** This dataset is our core behavioral predictor. It provides daily, anonymized data on population movement trends across categories like workplaces, transit\_stations, and residential. This directly measures the real-world behavior that facilitates or slows airborne disease transmission. This data is essential for capturing the subtle shifts in population mixing that precede an outbreak spike.

**Dataset 3: Government Policy Data**

* **Data Source:** Oxford COVID-19 Government Response Tracker (OxCGRT)
* **Data Format:** A single global CSV file.
* **Data Size:** Large, containing daily data for over 180 countries.
* **Why Chosen & Relevance:** This dataset provides the essential "context" that our model cannot learn from mobility or case data alone. It systematically tracks government interventions (like school closures, travel bans, and stay-at-home orders). Its most valuable feature is the **StringencyIndex**, a composite score from 0-100 that quantifies the overall strictness of government policy. By feeding this index into our model, we allow it to distinguish between a *voluntary* drop in mobility (e.g., due to public fear) and a *mandatory* one (e.g., due to a lockdown). This prevents the model from learning false relationships and is critical for making it robust.

#### 4. Data Analysis and Insights (Initial Exploration)

While the full data analysis is the next phase, our initial research reveals key expected patterns:

* **Key Insight (Mobility vs. Cases):** We expect to see a time-lagged correlation between mobility and new cases. For example, a sustained increase in the transit\_stations and workplaces mobility metrics will likely be followed by a rise in COVID-19 positivity rates 1-3 weeks later.
* **Key Insight (Policy as a Control):** The StringencyIndex will act as a powerful switch. When the index rises above a certain threshold (indicating a hard lockdown), we expect to see mobility in workplaces and retail drop sharply, followed by a subsequent, forced decline in case numbers.
* **Descriptive Statistics (Example):**
  + **Health Data:** We will plot the 7-day rolling average of the positive\_rate to identify the major "waves" (e.g., Delta, Omicron) in our chosen country.
  + **Mobility Data:** We will plot the workplaces\_percent\_change\_from\_baseline against the residential\_percent\_change\_from\_baseline. We expect to see a strong inverse correlation, as people who are not at work are at home.

#### 5. Conclusion

This data research phase has identified a robust, multi-dimensional set of data for our project. The key finding is that forecasting a complex event like the COVID-19 pandemic is impossible with any single data source.

By combining **Health Data (the target)**, **Mobility Data (the behavior)**, and **Policy Data (the context)**, we can build a predictive model that understands the complex interplay between the virus, the population, and the government's response. This integrated approach is essential for achieving our project's goal of creating a meaningful and reliable early warning system for airborne disease outbreaks.

#### 6. Citations

1. **Health Data:** Hasell, J., Mathieu, E., Beltekian, D. et al. "A cross-country database of COVID-19 testing." *Sci Data* 7, 345 (2020). Retrieved from: https://github.com/owid/covid-19-data
2. **Mobility Data:** Google. "Google COVID-19 Community Mobility Reports." (2022). Retrieved from: https://www.google.com/covid19/mobility/
3. **Policy Data:** Hale, T., Angrist, N., Goldszmidt, R. et al. "A global panel database of government responses to COVID-19." *Sci Data* 8, 111 (2021). Retrieved from: https://github.com/OxCGRT/covid-policy-tracker

### 

Part 3: Technology Review

#### **1. Introduction**

* **Importance & Relevance:** Accurate epidemic forecasting requires technologies capable of modeling complex, multi-source, and time-lagged data. This review is critical to justify the selection of advanced deep learning models (LSTMs, Transformers, etc.) over simpler methods. These tools are directly relevant to the project's goal of building a robust early warning system, aligning with SDG 3 (Good Health and Well-Being).

#### **2. Technology Overview**

This review covers the primary models and tools for the forecasting task.

* **Random Forest:** An "ensemble" machine learning model that works by building many decision trees and averaging their predictions. It is not a sequence model but is powerful for classification and regression with tabular data.
* **GRU (Gated Recurrent Unit):** A type of Recurrent Neural Network (RNN) and a modern alternative to LSTMs. It uses "gates" to manage a memory state, allowing it to learn patterns and dependencies over time.
* **LSTM (Long Short-Term Memory):** The classic, powerful RNN model. It uses a more complex cell structure with three gates (input, output, forget) to manage its memory, making it highly effective at learning very long-term dependencies.
* **TCN (Temporal Convolutional Network):** A model that adapts Convolutional Neural Networks (CNNs), typically used for images, to sequence data. It uses layers of "dilated" convolutions to look at patterns over long time periods simultaneously.
* **Transformer:** A state-of-the-art model that relies entirely on a "self-attention mechanism." Instead of processing data in order (like an LSTM), it looks at the entire sequence at once and determines which time steps are most important to each other, allowing it to capture complex, global relationships.
* **Hybrid Models (e.g., LSTM-Transformer):** A modern approach that combines models. For example, an LSTM is used to first process the sequence and find local patterns, and its output is then fed into a Transformer to find global relationships.
* **Implementation Tools:**
  + **Python:** The core programming language.
  + **Pandas:** A Python library essential for data loading, cleaning, merging (e.g., WHO data + Google data), and feature engineering (creating "lag" features).
  + **TensorFlow / PyTorch:** The deep learning frameworks used to build, train, and evaluate the GRU, LSTM, TCN, and Transformer models.

#### **3. Relevance to the Project**

* **Random Forest** is relevant as a powerful **baseline**. It will show how well a non-sequential model performs, helping to justify the need for more complex deep learning.
* **LSTM & GRU** are directly relevant because they are *designed* to model the project's core challenge: **time-lagged relationships**. They can learn that a change in mobility *today* affects case numbers *two weeks from now*.
* **TCN & Transformer** are relevant for finding **more complex, non-obvious patterns** that the other models might miss. The Transformer, in particular, can see how mobility on "day 1" and "day 10" *jointly* affect "day 20," without being limited by the sequential steps in between.
* **Hybrid Models** are relevant as they represent the **current state-of-the-art** for this exact problem, combining the strengths of both LSTMs and Transformers.

#### **4. Comparison and Evaluation**

* **Random Forest**
  + **Ease of Use/Cost:** Very easy and fast to train. Computationally cheap.
  + **Scalability:** Scales well for data size, but not for sequence length.
  + **Performance:** Strong baseline. Excellent interpretability (we can see which features are most important). Its main weakness is that it struggles with long-term temporal dependencies unless features are manually and perfectly engineered.
* **LSTM (Long Short-Term Memory)**
  + **Ease of Use/Cost:** More complex and computationally expensive to train than Random Forest. Can be slow.
  + **Scalability:** Scales well to very long sequences.
  + **Performance:** Excellent at capturing long-term, time-lagged effects. It is a proven, reliable standard for this type of forecasting.
* **GRU (Gated Recurrent Unit)**
  + **Ease of Use/Cost:** Less complex and computationally *faster* to train than an LSTM because it has a simpler gate structure.
  + **Scalability:** Similar to LSTM.
  + **Performance:** Often achieves performance on par with LSTMs. It is a highly efficient and effective alternative.
* **TCN (Temporal Convolutional Network)**
  + **Ease of Use/Cost:** Can be faster to train than RNNs (LSTMs/GRUs) because its convolutional structure can be parallelized.
  + **Scalability:** Scales very well and can have a large "receptive field" to see long-ago patterns.
  + **Performance:** Very strong performance, especially on long sequences. It is an excellent, modern alternative to RNNs.
* **Transformer**
  + **Ease of Use/Cost:** The most complex and computationally expensive model. It requires significant data and careful tuning ("hyperparameter tuning").
  + **Scalability:** Scales extremely well to massive datasets, which is its main advantage.
  + **Performance:** Often state-of-the-art, as it can capture complex global interactions that all other models miss. However, it can overfit smaller datasets.
* **Hybrid (e.g., LSTM-Transformer)**
  + **Ease of Use/Cost:** The most complex of all, as it involves tuning two separate models.
  + **Scalability:** Similar to the Transformer.
  + **Performance:** Has the potential to be the best-performing model, as it synergistically combines the strengths of both approaches.

#### **5. Use Cases and Examples**

* **Random Forest:** Often used as a baseline model in forecasting studies, such as in **Ayyoubzadeh et al. (2020)**, to prove the value of more complex models.
* **LSTM / GRU:** These are the most common models in modern epidemic forecasting literature, including papers that use mobility data.
* **TCN:** Widely used in other complex time-series tasks like audio processing and medical signal (e.g., EKG) analysis, as shown by **Bai et al. (2018)**.
* **Transformer / Hybrid:** This represents the cutting edge. **Jiao et al. (2025)** specifically used a hybrid LSTM-Transformer model to forecast COVID-19 in Japan, finding it *outperformed* standalone models. This provides a direct, state-of-the-art precedent for this project.

#### **6. Identify Gaps and Research Opportunities**

* **Explainability:** The biggest gap for advanced models (Transformer, TCN, Hybrid) is a lack of interpretability. It's hard to know *why* they made a certain prediction. A research opportunity is to apply tools like SHAP or LIME to try and "look inside the box."
* **Comparison:** While many papers use one of these models, few comprehensively compare all of them (Random Forest vs. GRU vs. TCN vs. Transformer) on the *same* public health dataset. This project can fill that gap.
* **Customization:** There is an opportunity to customize these models for real-time surveillance.

#### **7. Conclusion**

* **Key Takeaways:** This review shows a clear progression from simple models to highly complex, state-of-the-art hybrid architectures.
* **Importance of Chosen Tools:** The chosen technologies (Random Forest, GRU, LSTM, TCN, Transformer, and Hybrids) are not redundant; they each serve a specific purpose: establishing a baseline (Random Forest), applying the reliable standard (LSTM/GRU), and pushing for state-of-the-art accuracy (TCN/Transformer/Hybrid).
* **Benefit to Project:** This multi-model approach, justified by recent literature like **Jiao et al. (2025)**, creates a robust, comparative study. It moves beyond just *building a forecast* and contributes by *evaluating which forecasting method is best* for this specific public health challenge.

#### **8. Citations**

* Ayyoubzadeh, S. M., et al. (2020). Predicting COVID-19 incidence through Google Trends and mobility data. *Health Information Science and Systems*.
* Bai, S., et al. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271*.
* Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.
* Jiao, S., et al. (2025). Spatio-temporal epidemic forecasting using mobility data and deep learning. *Scientific Reports*.
* Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*.