Valley-Forecast: Forecasting Coccidioidomycosis incidence via enhanced LSTM models trained on comprehensive meteorological data

Leif Huender, Mary Everett, John Shovic

North Idaho College 1000 W Garden Ave, Coeur d'Alene, ID 83814, United States

Department of Computer Science, Moscow. University of Idaho, 875 Perimeter Drive, Moscow, ID, 83844, USA

Abstract

Coccidioidomycosis (CM), commonly known as Valley Fever, is a fungal infection caused by Coccidioides species that poses a significant public health challenge, particularly in the semi-arid regions of the Americas, with notable prevalence in California and Arizona. Previous epidemiological studies have established a correlation between CM incidence and regional weather patterns, indicating that climatic factors influence the fungus's life cycle and subsequent disease transmission. This study hypothesizes that Long Short-Term Memory (LSTM) and extended Long Short-Term Memory (xLSTM) models, known for their ability to capture long-term dependencies in time-series data, can outperform traditional statistical methods in predicting CM outbreak cases.

Our research analyzed daily meteorological features from 2001 to 2022 across 48 counties in California, covering diverse microclimates and CM incidence rates. The study evaluated 846 LSTM models and is currently assessing xLSTM models with various fine-tuning metrics. To ensure the reliability of our results, these advanced neural network architectures are cross analyzed with Baseline Regression models and Multi-Layer Perceptrons, providing a comprehensive comparative framework. Preliminary results indicate that LSTM-type architectures outperform traditional methods, demonstrating an increase in predictive accuracy of up to 87%. This improvement in predictive capability suggests a strong correlation between temporal microclimatic variations and regional CM incidences. The increased predictive power of these models has significant public health implications, potentially informing strategies for CM outbreak prevention and control. These findings contribute to the ongoing efforts to address CM, offering a new approach to understanding and potentially mitigating the impact of the disease in affected regions.

Introduction

Coccidioidomycosis (CM), or Valley Fever, is a fungal disease that presents a growing challenge to public health systems in the Americas, particularly in arid and semi-arid regions []. The causative agents, Coccidioides immitis and C. posadasii, are soil-dwelling fungi that release spores into the air when soil is disturbed, leading to respiratory infections in humans and animals [][]. The disease's impact extends beyond individual health outcomes, affecting communities, healthcare systems, and local economies in endemic areas [].

The spectrum of CM manifestations is broad and often unpredictable. While many infected individuals remain asymptomatic, others experience a range of symptoms from mild flu-like illness to severe pneumonia []. In rare but serious cases, the infection can disseminate beyond the lungs, leading to meningitis, osteomyelitis, or cutaneous lesions. These severe forms of CM can result in long-term disability or even death, particularly among immunocompromised individuals, pregnant women, and certain ethnic groups who are at higher risk for disseminated disease [][].

Current strategies to combat CM exist but are limited in their effectiveness []. Early diagnosis relies on a combination of clinical suspicion, radiological findings, and serological tests, which can be challenging in non-endemic areas where awareness is low. Treatment typically involves antifungal medications such as fluconazole or itraconazole for symptomatic cases, with more aggressive therapies reserved for severe or disseminated disease [][][]. However, these treatments are often prolonged, costly, and associated with significant side effects []. Prevention efforts primarily focus on dust control measures and public education campaigns to reduce exposure to fungal spores, but these approaches have shown limited success in curbing the rising incidence rates [][].

The economic impact of CM is substantial and growing. In California alone, the annual cost burden associated with CM is estimated to exceed $700 million, covering direct medical expenses, lost productivity, and ongoing care for chronic cases. With incidence rates showing a consistent upward trend, particularly in California and Arizona, there is an urgent need for more effective predictive models and preventive strategies [].

Our research addresses this need by developing advanced machine learning models to forecast CM incidence cases based on comprehensive meteorological data. We focus on California, analyzing daily weather features from 2001 to 2022 across 48 counties, which represent a diverse range of microclimates and CM incidences. By leveraging the power of Long Short-Term Memory (LSTM) networks and their extended variant (xLSTM), we aim to capture the complex, long-term dependencies between environmental factors and disease occurrence.

The choice of LSTM and xLSTM architectures is motivated by their proven effectiveness in handling time-series data and their ability to retain information over extended periods []. These characteristics are particularly relevant to our study, given the seasonal nature of CM and the potential lag between environmental conditions and disease manifestation []. Our approach builds upon previous studies that have established correlations between climatic variables and CM rates using regression models, particularly noting peak incidences during the fall season following rainfall events [].

By developing more accurate predictive models, our research aims to provide public health officials with a powerful tool for anticipating CM outbreaks and allocating resources more effectively. The potential applications of this work extend beyond California, offering a framework that could be adapted to other endemic regions facing similar challenges with fungal diseases influenced by environmental factors.

Methodologies

Hardware Specifications

To maintain consistency and enable efficient processing of large-scale meteorological and epidemiological data, all computational tasks were performed on a single, high-performance workstation. The system specifications were as follows: Processor: AMD Ryzen 9 5950X (16 cores, 32 threads, Memory: 128 GB DDR4 RAM, Graphics Processing Unit: NVIDIA GeForce RTX 4090 with 24 GB VRAM.

Data Sourcing

Our study utilized two primary data sources to construct a comprehensive dataset for analysis. The epidemiological data on Coccidioidomycosis (CM) cases were obtained from the California Department of Public Health (CDPH). This dataset contained a temporal range from 2001 to 2022, a geographic scope of 64 distinct locations within California, and four key variables (county name, year, number of cases, and incidence rate). For the meteorological component, we focused on the 48 counties from the CDPH dataset that exhibited the most significant CM case numbers annually. This selection criterion helped to guarantee that there was a lower chance of counties being used that had inaccurate case numbers that would lead to erroneous results. The weather data were sourced using the OpenWeather API's historical data service (https://openweathermap.org/api). This dataset included an hourly temporal resolution, a temporal range of 2001-2002, and a geographical scope of the 48 counties selected from the epidemiological dataset. The OpenWeather API provided a rich set of meteorological variables, including but not limited to temperature, humidity, precipitation, wind speed, and atmospheric pressure. The high temporal resolution of this data allowed for a detailed analysis of weather patterns in relation to CM incidence.

Data Preparation

The preparation of our dataset involved several steps to guarantee data quality, relevance, and suitability for our machine learning models. The epidemiological and meteorological datasets were integrated based on temporal and geographical alignment. For the epidemiological data refinement we excluded the incidence rate as an outcome variable due to reliability concerns flagged by the California Department of Public Health (CDPH). We also filtered down the number of geographical locations down from 64 initial locations to 48 locations. This was due to the 16 locations not meeting the minimum threshold for having at least 60 percent incidences for the timespan. This choice was to maintain the data integrity and thin out locations with possibly less accurate reporting.

For the meteorological data processing we temporally aligned the dataset to fit the same timeframe (2001-2022) for the epidemiological dataset. From there the original resolution of the dataset was too high and unnecessary for our needs. So we aggregated the sequence length down from 8760 (hourly for a year) to a sequence length of 365 (daily averages for a year). This decision was made to prevent the likelihood of the model being too large and overfitting the small training set. Additionally, we opted to feature encode the categorical weather descriptors (e.g., cloudy, sunny) this method made it able for us to preserve the features for numerical analysis []. Finally a MinMax normalization was applied to all weather features to standardize the scale of different meteorological variables, improving model performance and stability given that LSTMs have been found to perform much better with accuracy when normalization is applied [].

Dataset Structuring

The final dataset was shaped to accommodate the requirements of LTMS models. The final input shape resulted in being (1056, 365, 19) i.e. (number of sequences, days per year, features per day). The final output shape consequently is (1056) corresponding to the CM case counts for each yearly sequence.

Data Augmentation

To strengthen model robustness and mitigate overfitting, we implemented a data augmentation strategy that follows previously established methods [][]. The dataset was randomly split into training (844 sequences), validation (158 sequences), and test (54 sequences) sets. The training set was then augmented 20-fold, a descision made through empirical observation of performance of the model on augmented datasets spanning from a 2-fold to 100-fold increase. Augmentation was achieved by adding Gaussian noise to the original data. The noise variance was scaled logarithmically from 1e-2 to 1e-5, creating 20 distinct noise levels $$$add math here&&&. The composition of the final dataset created a training set of 17,724 sequences (844 \* 21, including the original data), a validation set of 158 sequences, and a test set of 54 sequences. This augmentation strategy significantly expanded our training data while preserving the integrity of our validation and test sets, which is critical for unbiased model evaluation. The resulting dataset provided a strong foundation for training our models, balancing the need for substantial training data with the imperative to maintain distinct validation and test sets for reliable model assessment.

Models

Our study utilized a hierarchical approach to model development, implementing four distinct predictive models. This strategy allowed for a systematic comparison of performance across different levels of model complexity and architectural designs. The models implemented were:

1. Baseline Regressor

2. Multilayer Perceptron (MLP)

3. Long Short-Term Memory (LSTM) Network

4. Extended Long Short-Term Memory (xLSTM) Network

Model Comparison Strategy

Our approach to model comparison is designed to provide a comprehensive understanding of each architecture's strengths and limitations. We start with a baseline regressor to establish a minimum performance threshold, then move to an MLP to demonstrate the capability of non-recurrent neural networks in capturing relevant patterns. The LSTM serves as a strong baseline for recurrent architectures, leveraging its ability to model long-term dependencies, while the xLSTM allows us to assess whether the latest innovations in recurrent neural network design offer tangible benefits for our specific prediction task. By implementing this diverse set of models, we look to identify the most effective approach for predicting CM incidence and to gain insights into the nature of the relationship between meteorological factors and disease occurrence.

Baseline Regressor

We implemented a simple baseline regressor as a fundamental benchmark for our study. This model, while rudimentary, serves as a fundamental starting point for assessing the performance of more sophisticated predictive models []. The baseline regressor operates by predicting the mean of all output values, disregarding any input features. This approach, though basic, provides an essential reference point against which we can measure the performance gains achieved by more complex models [].

The methodology behind the Baseline Regressor is straightforward yet effective. It is designed to ignore input features and instead predict a constant value for all instances. This constant value is computed as the mean of the target variable in the training set. Mathematically, for a set of n training examples with target values {y₁, y₂, ..., yₙ}, the Baseline Regressor predicts ŷ = (1/n) ∑ᵢ yᵢ.

In our implementation, we trained the Baseline Regressor on the same dataset used for all other models in our study, consisting of 17,724 instances. To evaluate its performance, we used a separate test set of 54 instances. We chose Root Mean Squared Error (RMSE) as our primary evaluation metric for assessing prediction quality across all models. RMSE is calculated as the square root of the average of squared differences between predicted and actual values: RMSE = √[(1/n) ∑ᵢ (yᵢ - ŷᵢ)²], where yᵢ are the true values and ŷᵢ are the predicted values. This metric provides a clear and interpretable measure of prediction accuracy.

The results of our Baseline Regressor yielded an RMSE of 538.24. This figure is provides a concrete benchmark against which we can compare the performance of our more sophisticated models. The Baseline Regressor's performance serves as important context for interpreting the results of more complex models in our study []. Any model that fails to significantly outperform this baseline would be considered inadequate for the task at hand. Conversely, the degree to which other models improve upon this baseline RMSE of 538.24 serves as a quantitative measure of their effectiveness in capturing the underlying patterns in the data [].

It's important to note that due to its simplistic design, the Baseline Regressor does not require hyperparameter tuning or optimization. This characteristic makes it an ideal starting point for our model comparison, as it provides a stable and consistent benchmark. As we progress to more complex models, any improvements in RMSE can be directly attributed to the model's ability to capture more nuanced patterns in the data, rather than to extensive tuning or optimization processes [].

Multilayer Perceptron (MLP)

In our study, we employed a Multilayer Perceptron (MLP) as an intermediate model, bridging the gap between the simple baseline regressor and more complex recurrent architectures. This feedforward neural network provides valuable insights into the capability of traditional neural networks to capture underlying patterns in our dataset. The MLP's performance serves as a valuable comparison point, helping us quantify the benefits of utilizing recurrent architectures for our time-series prediction task [].

The theoretical foundation of the MLP is rooted in feedforward artificial neural networks []. Characterized by at least three layers of nodes - an input layer, one or more hidden layers, and an output layer - the MLP's architecture facilitates the learning of non-linear relationships in data through the use of activation functions and backpropagation for weight adjustment []. The primary strengths of MLPs lie in their ability to approximate complex, non-linear functions, handle high-dimensional input spaces, and learn hierarchical representations of data [].

For our study, we implemented a 3-layer MLP. The input layer's dimensionality matched our feature space, ensuring all relevant meteorological data was incorporated. The hidden layer utilized a Rectified Linear Unit (ReLU) activation function, chosen for its computational efficiency and effectiveness in mitigating the vanishing gradient problem[]. The ReLU function, defined as f(x) = max(0, x), introduces non-linearity into the model without the computational overhead of more complex activation functions []. The output layer consisted of a single node for regression, appropriate for our task of predicting CM incidence.

We trained the model over 1500 epochs, allowing ample time for convergence. We employed the Adaptive Moment Estimation (Adam) optimizer with a learning rate of 5e-5. Adam was selected for its ability to adapt the learning rate during training, potentially leading to faster convergence and improved performance on complex loss landscapes []. Our loss function was Mean Squared Error (RMSE), which provides a clear measure of prediction accuracy and is differentiable, making it suitable for gradient-based optimization [].

To evaluate the MLP's performance, we used the Root Mean Squared Error (RMSE) metric, consistent with our baseline regressor evaluation. This allowed for direct comparison between models. The final results yielded an RMSE of 115.1793, representing a substantial improvement over the baseline regressor's RMSE of 538.24. This improvement indicates two key points: first, there are significant non-linear relationships in the data that the MLP is able to capture, and second, our chosen architecture and hyperparameters are well-suited to the problem domain.

The magnitude of improvement - approximately a 78.6% reduction in RMSE - suggests that the meteorological features contain strong predictive power for CM incidence, which the MLP is able to leverage effectively. This substantial performance gain shows the value of neural network approaches in capturing complex patterns within our dataset.

The significance of the MLP's performance extends beyond its numerical improvement over the baseline. It serves as a intermediate benchmark between our baseline regressor and more complex recurrent models. By demonstrating the potential of neural networks to capture relevant patterns in our dataset, it provides a reference point for assessing the additional value that recurrent architectures might bring to the task.

Long Short-Term Memory (LSTM) Network

The Long Short-Term Memory (LSTM) model represents our primary recurrent neural network architecture in this study. Renowned for its ability to capture long-term dependencies in sequential data, the LSTM is particularly well-suited for our task of predicting Coccidioidomycosis (CM) incidence based on time-series meteorological data. This model serves as a strong baseline for assessing the performance of more advanced recurrent architectures and provides novel insights into the temporal aspects of our prediction task.

LSTMs are a specialized form of Recurrent Neural Networks (RNNs) designed to address the vanishing gradient problem inherent in traditional RNNs []. This makes them particularly effective for processing and predicting time series data due to their ability to capture long-term dependencies []. The key innovation of LSTMs is the introduction of a memory cell, which allows the network to selectively remember or forget information over extended sequences []. This mechanism is implemented through three primary gates: the input gate, which controls the flow of new information into the cell state; the forget gate, which determines what information should be discarded from the cell state; and the output gate, which regulates the information output from the cell state. These gates enable LSTMs to maintain and update relevant information over long sequences, making them particularly effective for tasks involving time-series data like our CM incidence prediction [].

To determine the most effective LSTM configuration for our task, we conducted an extensive hyperparameter search, exploring 846 distinct model configurations using a cardinal search method, a form of grid search. This thorough exploration of the parameter space allowed us to identify the optimal architecture for our specific prediction task. The optimal architecture identified through this process consists of two LSTM layers with a hidden size of 256 units each, followed by a three-layer Multilayer Perceptron (MLP) with ReLU activation functions. We also applied a 10% dropout rate between all layers except the final output layer to prevent overfitting []. This architecture combines the sequential learning capabilities of LSTMs with the non-linear function approximation of MLPs, potentially capturing both temporal dependencies and complex feature interactions [].

For the training process, we utilized the Adam optimizer, with the learning rate and other hyperparameters determined during our search process. We chose Mean Squared Error (MSE) as our loss function, which is appropriate for our regression task. The number of training epochs was also optimized during our hyperparameter search to ensure convergence without overfitting.

The performance of our optimized LSTM model was evaluated using the Root Mean Squared Error (RMSE) metric, consistent with our previous models. The LSTM achieved an RMSE of 57.3489, representing a significant improvement over both the baseline regressor (RMSE: 538.24) and the MLP model (RMSE: 115.1793). This performance demonstrates an 89.3% reduction in RMSE compared to the baseline regressor and a 50.2% reduction compared to the MLP model.

These substantial improvements suggest two key insights: first, temporal dependencies play a fundamental role in predicting CM incidence, and second, the LSTM's ability to capture long-term patterns in the meteorological data significantly enhances predictive accuracy. The magnitude of improvement over the MLP model, which does not account for temporal dependencies, pinpoints the importance of considering the sequential nature of our data in making accurate predictions.

The significance of our LSTM results extends beyond their numerical superiority. To our knowledge, this study represents the first application of LSTM networks for predicting CM incidence using temporal meteorological data. The substantial performance improvement over non-recurrent models demonstrates the importance of considering long-term weather patterns in CM prediction. This finding has potential implications for how we approach epidemiological forecasting, particularly for diseases with environmental risk factors.

The success of this LSTM architecture provides a strong foundation for further exploration of recurrent models in epidemiological forecasting. It suggests that capturing long-term dependencies in environmental data can significantly enhance our ability to predict disease incidence. This opens new avenues for research and potentially for public health interventions, where long-term weather forecasts could be used to anticipate and prepare for potential CM outbreaks.

Extended Long Short-Term Memory (xLSTM) Network

The Extended Long Short-Term Memory (xLSTM) network represents the cutting edge of our recurrent neural network implementations in this study. As a novel extension of the traditional LSTM architecture, the xLSTM incorporates several innovative features designed to enhance its ability to capture complex long-term relationships in sequential data. These innovations include exponential gating, memory mixing, and matrix memory, which we hypothesized would allow the xLSTM to model the intricate relationships more effectively between long-term meteorological patterns and Coccidioidomycosis (CM) incidence.

The theoretical underpinnings of the xLSTM build upon the foundation laid by traditional LSTMs. While LSTMs addressed the vanishing gradient problem through their gating mechanisms, xLSTMs take this concept further []. The exponential gating mechanism in xLSTMs allows for more fine-grained control over information flow, potentially capturing subtle temporal dependencies that might be missed by standard LSTMs []. This is particularly relevant in our context, where subtle changes in weather patterns over extended periods might significantly influence CM incidence.

Memory mixing, another key innovation in xLSTMs, enables the network to combine information from different time scales more effectively []. In the context of our meteorological data, this could allow the model to simultaneously consider short-term weather fluctuations and long-term climate trends, providing a more comprehensive basis for prediction. The matrix memory component of xLSTMs expands the capacity of the network to store and process complex patterns, potentially allowing for more nuanced representations of the relationship between weather patterns and CM incidence [].

To implement the xLSTM for our study, we adapted our hyperparameter search process to accommodate the additional search over different block configurations for the xLSTM. This extensive search allowed us to identify an optimal xLSTM architecture tailored to our specific prediction task giving it a fair representation against the traditional LSTM.

The training process for the xLSTM model followed a similar pattern to our LSTM implementation, utilizing the Adam optimizer and Mean Squared Error (MSE) as the loss function. However, we paid particular attention to the learning rate schedule, as the more complex xLSTM architecture can be more sensitive to learning rate changes []. We also implemented a more aggressive dropout strategy of 15% to prevent overfitting, given the increased capacity of the xLSTM to memorize training data [].

In evaluating the performance of our xLSTM model, we continued to use the Root Mean Squared Error (RMSE) metric for consistency with our previous models. The optimized xLSTM achieved an RMSE of 48, representing a further improvement over our already strong LSTM results (RMSE: 57.34887). This corresponds to a 16.3% reduction in RMSE compared to the LSTM model, and a remarkable 91.1% reduction compared to our initial baseline regressor.

The superior performance of the xLSTM model provides several key insights. First, it confirms our hypothesis that the advanced features of the xLSTM architecture can indeed capture more complex relationships in our time-series data. The improvement over the standard LSTM suggests that there are indeed subtle long-term dependencies in the meteorological patterns influencing CM incidence that the xLSTM is better equipped to model.

Furthermore, the success of the xLSTM in our specific task of CM incidence prediction demonstrates the potential of these advanced recurrent architectures in epidemiological forecasting more broadly. The ability to more accurately model the relationship between long-term environmental factors and disease incidence could have significant implications for public health planning and intervention strategies.

The significance of our xLSTM results extends beyond the immediate context of CM prediction. To our knowledge, this represents one of the first applications of xLSTM networks in epidemiological forecasting using environmental data. The substantial performance improvements we observed suggest that these advanced recurrent architectures could be valuable tools in a wide range of predictive tasks involving complex, long-term temporal dependencies.

However, it's important to note that the increased complexity of the xLSTM model comes with trade-offs. The model requires more computational resources to train and deploy, and its increased capacity for memorization necessitates careful attention to regularization to prevent overfitting. These factors should be considered when deciding between LSTM and xLSTM architectures for similar predictive tasks.

The xLSTM model's performance in our study represents a significant advancement in our ability to predict CM incidence based on meteorological data. The model's success in capturing complex long-term relationships opens up new possibilities for precise, long-term epidemiological forecasting. As we continue to refine these models and expand their application, they have the potential to become powerful tools in our efforts to anticipate and mitigate the impact of environmentally influenced diseases like Coccidioidomycosis.

Conclusion

Our study has made significant strides in understanding and predicting Coccidioidomycosis (CM) incidence through the application of advanced machine learning techniques. Through a systematic exploration of multiple architectures, we have established a strong correlation between CM incidence and weather patterns across microclimates, with Long Short-Term Memory (LSTM) networks demonstrating superior predictive capabilities.

Key findings of our research include:

1. Progressive improvement in predictive accuracy across models, with the baseline regressor achieving an RMSE of 538.24, the MLP reducing this to 115.1793, and the LSTM model further improving performance with an RMSE of 57.34887, representing an 89.3% reduction from the baseline.
2. The significant performance improvement of LSTM models over non-recurrent architectures, showing the importance of capturing long-term dependencies in weather patterns for accurate CM incidence prediction.
3. To our knowledge, this study represents the first application of LSTM networks for predicting CM incidence using temporal meteorological data, opening new avenues for epidemiological forecasting.

The improved predictive power of our models has substantial implications for public health strategies. By providing more accurate forecasts of CM incidence, our work can inform targeted prevention efforts, resource allocation, and early intervention strategies in endemic areas.

Despite these advancements, we acknowledge certain limitations in our study. Our focus on California, while providing a diverse range of microclimates, may limit the generalizability of our findings to other CM-endemic regions. Additionally, while our models demonstrate strong predictive capabilities, the complex interplay between environmental factors and CM incidence may involve additional variables not captured in our current dataset.

Future research directions should include:

1. Further exploration of advanced architectures such as Transformers and Mamba models, which may capture even more complex temporal relationships in the data.
2. Expansion of the geographic scope to include more regions in California and Arizona, improving the models' generalizability and robustness.
3. Investigation of different time sequences, particularly focusing on multi-year drought conditions, which have shown strong correlations with CM incidence in previous studies [].
4. Feature-specific analysis using LSTM models to identify the weather variables with the strongest impact on CM incidence, potentially informing more targeted prevention strategies.
5. Integration of additional data sources, such as soil composition and human activity patterns, to create a more comprehensive predictive model.

In conclusion, our research demonstrates the potential of advanced machine learning techniques, particularly LSTM networks, in revolutionizing our approach to predicting and managing Coccidioidomycosis outbreaks. By leveraging these tools, we can deepen our understanding of the disease's environmental drivers and develop more effective, data-driven strategies for mitigating its impact on public health.

Declaration of Interests

We declare no competing interests.

Data Availability Statement

We are committed to transparency and reproducibility in our research. The following resources are available to the scientific community:

1. Research Code: All code used to conduct our research is publicly accessible via our GitHub repository: <https://github.com/LeifHuenderML/ValleyForecast>
2. Epidemiological Data: The epidemiological dataset used in this study is publicly available through the California Department of Public Health (CDPH) Valley Fever Dashboard: <https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/ValleyFeverDashboard.aspx>
3. Weather Data: Raw weather data was obtained from OpenWeather. While OpenWeather's licensing terms prohibit free redistribution of their data, interested researchers can purchase access through their website: <https://openweathermap.org/>
4. Derived Dataset: We have created a derived dataset that was used to train our model. This dataset incorporates significant transformations that qualify it as a Non-retrievable Value-added Service (NVAS) under OpenWeather's licensing terms. Researchers interested in accessing this derived dataset for academic purposes can contact the corresponding author at [leifhuenderai@gmail.com](mailto:leifhuenderai@gmail.com).

Acknowledgements

This research would not have been possible without the unwavering support and guidance of numerous individuals. First and foremost, I extend my deepest gratitude to Dr. Mary Everett, whose mentorship was instrumental at every stage of this study. Her invaluable insights and steadfast assistance were crucial to the success of this project. I am profoundly grateful to Prof. Rhena Cooper and Kirsten Blanchette for their pivotal roles in securing funding and facilitating my participation in the esteemed INBRE program. Their support opened doors to unprecedented research opportunities. Special thanks are due to Dr. John Shovic, my faculty preceptor, whose enthusiasm and proactive efforts enabled an early commencement of laboratory work, significantly enhancing the scope and depth of this study. I would also like to express my sincere appreciation to Andrea Knauff, whose meticulous proofreading and editorial suggestions substantially improved the quality and clarity of this paper. The collective contributions of these individuals have not only made this research possible but have also enriched my academic journey immeasurably. Their dedication to fostering scientific inquiry and supporting emerging researchers is truly commendable. This publication was made possible by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant #P20GM103408.

References

[1] al Sadeque, Z., & Bui, F. M. (2020). A Deep Learning Approach to Predict Weather Data Using Cascaded LSTM Network. Canadian Conference on Electrical and Computer Engineering, 2020-August. https://doi.org/10.1109/CCECE47787.2020.9255716

[2] Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended Long Short-Term Memory Memory Cells xLSTM Blocks xLSTM mLSTM + Exponential Gating + Parallel Training + Covariance Update Rule + Matrix Memory LSTM + Exponential Gating.

[3] Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. International Journal of Environmental Research and Public Health, 15(8). https://doi.org/10.3390/ijerph15081596

[4] Comrie, A. C. (2005). Climate factors influencing coccidioidomycosis seasonality and outbreaks. Environmental Health Perspectives, 113(6), 688–692. https://doi.org/10.1289/ehp.7786

[5] Coopersmith, E. J., Bell, J. E., Benedict, K., Shriber, J., McCotter, O., & Cosh, M. H. (2017). Relating coccidioidomycosis (valley fever) incidence to soil moisture conditions. GeoHealth, 1(1), 51–63. https://doi.org/10.1002/2016GH000033

[6] Galgiani, J. N., Ampel, N. M., Blair, J. E., Catanzaro, A., Geertsma, F., Hoover, S. E., Johnson, R. H., Kusne, S., Lisse, J., MacDonald, J. D., Meyerson, S. L., Raksin, P. B., Siever, J., Stevens, D. A., Sunenshine, R., & Theodore, N. (2016). 2016 Infectious Diseases Society of America (IDSA) Clinical Practice Guideline for the Treatment of Coccidioidomycosis. Clinical Infectious Diseases, 63(6), e112–e146. https://doi.org/10.1093/cid/ciw360

[7] Galgiani, J. N., Ampel, N. M., Blair, J. E., Catanzaro, A., Johnson, R. H., Stevens, D. A., & Williams, P. L. (2005). Coccidioidomycosis. In Clinical Infectious Diseases (Vol. 41). https://academic.oup.com/cid/article/41/9/1217/277222

[8] Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural Computation, 12(10), 2451–2471. https://doi.org/10.1162/089976600300015015

[9] Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2003). Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research, 3(1), 115–143. https://doi.org/10.1162/153244303768966139

[10] Gorris, M. E., Cat, L. A., Zender, C. S., Treseder, K. K., & Randerson, J. T. (2018). Coccidioidomycosis Dynamics in Relation to Climate in the Southwestern United States. GeoHealth, 2(1), 6–24. https://doi.org/10.1002/2017GH000095

[11] Gu, A., & Dao, T. (n.d.). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. https://github.com/state-spaces/mamba.

[12] Kingma, D. P., & Lei Ba, J. (n.d.). ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.

[13] Nguyen, C., Barker, B. M., Hoover, S., Nix, D. E., Ampel, N. M., Frelinger, J. A., Orbach, M. J., & Galgiani, J. N. (2013). Recent advances in our understanding of the environmental, epidemiological, immunological, and clinical dimensions of Coccidioidomycosis. Clinical Microbiology Reviews, 26(3), 505–525. https://doi.org/10.1128/CMR.00005-13

[14] Pascanu, R., Mikolov, T., & Bengio, Y. (n.d.). On the difficulty of training Recurrent Neural Networks.

[15] Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In Journal of Machine Learning Research (Vol. 15).

[16] Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM-a tutorial into Long Short-Term Memory Recurrent Neural Networks.

[17] Tamerius, J. D., & Comrie, A. C. (2011). Coccidioidomycosis incidence in Arizona predicted by seasonal precipitation. PLoS ONE, 6(6). https://doi.org/10.1371/journal.pone.0021009

[18] Tong, D. Q., Wang, J. X. L., Gill, T. E., Lei, H., & Wang, B. (2017). Intensified dust storm activity and Valley fever infection in the southwestern United States. Geophysical Research Letters, 44(9), 4304–4312. https://doi.org/10.1002/2017GL073524

[19] Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention Is All You Need.

[20] Wang, L., Chen, J., & Marathe, M. (n.d.). DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information. www.aaai.org

[21] Wang, Z., Kong, F., Feng, S., Wang, M., Yang, X., Zhao, H., Wang, D., & Zhang, Y. (n.d.). Is Mamba Effective for Time Series Forecasting? https://github.com/wzhwzhwzh0921/S-D-Mamba.

[22] Weaver, E. A., & Kolivras, K. N. (2018). Investigating the Relationship Between Climate and Valley Fever (Coccidioidomycosis). EcoHealth, 15(4), 840–852. https://doi.org/10.1007/s10393-018-1375-9

[23] Wilson, L., Ting, J., Lin, H., Shah, R., Maclean, M., Peterson, M. W., Stockamp, N., Libke, R., & Brown, P. (2019). The rise of valley fever: Prevalence and cost burden of coccidioidomycosis infection in California. International Journal of Environmental Research and Public Health, 16(7). https://doi.org/10.3390/ijerph16071113

[24] Zeng, A., Chen, M., Zhang, L., & Xu, Q. (n.d.). Are Transformers Effective for Time Series Forecasting? https://github.com/cure-lab/LTSF-