
BioCast AI: Neural Forecasting of Global Disease Outbreaks¹

Muha Sunag Igor
mmad Parasu Mizin
Usman Nagsh
Saarla Saarla Online
nd nd Attend
Univer Univer ant
sity sity

AI AI
Safety Safety
Saarla Saarla
nd nd

With
Apart Research

Abstract

•

BioCast AI is a deep learning powered system designed to forecast global disease outbreaks using neural sequence models and advanced machine learning ensembles. Our approach integrates three architectures RNN LSTM and Bidirectional LSTM to capture long term temporal patterns in more than forty years of global disease data. In parallel we developed a feature rich ensemble model using XGBoost LightGBM Random Forests and Gradient Boosting to enhance predictive accuracy and stability.

We trained models on cleaned and imputed WHO style datasets covering 1980 to 2024 generating country specific and disease specific predictions for 2025. BioCast AI also includes a full

¹ Research conducted at the Defensive Acceleration Hackathon, 2025

analytical pipeline with PCA correlation analysis, temporal visualization clustering and dashboards to support epidemiological decision making.

The system achieved strong performance across models with ensemble methods outperforming individual networks in R^2 and RMSE scores. Outbreak risk levels were derived using predicted case trajectories, trend ratios and statistical confidence intervals. BioCast AI demonstrates how deep learning and ML ensembles can support early warning systems, public health forecasting and global disease surveillance.

Keywords: Disease forecasting, LSTM, time series modeling, outbreak prediction, epidemiological, AI, global health analytics

1. Introduction

BioCast AI addresses the growing global need for reliable scalable and data driven disease forecasting. Infectious diseases spread rapidly and real time predictive systems are essential for preparedness. With decades of historical case data available but underutilized machine learning offers a transformative path forward.

Our project leverages neural networks and ensemble models to predict outbreaks by learning subtle temporal structures from longitudinal case counts spanning 1980 to 2024. Compared with classical forecasting methods neural sequence models capture long range dependencies, seasonal effects and sudden spikes more effectively. Ensemble based models incorporate engineered statistical features improving stability and generalizability across diseases and countries.

BioCast AI contributes to epidemiological intelligence by unifying prediction analysis visualization and risk classification into one system. It showcases how computational tools can inform policy decisions, resource allocation and early warnings for high risk regions. The threat model focuses on natural infectious diseases and cross regional outbreak escalation patterns. By making prediction pipelines transparent and reproducible this work supports the mission of global health resilience.

Research Questions

- 1 Can neural networks reliably predict future disease case counts based on historical patterns
- 2 Does combining deep learning with feature driven ensembles improve forecasting accuracy
- 3 Can predictions be translated into interpretable outbreak risk scores for policymakers

Hypothesis

Ensembling neural sequence models with feature rich machine learning regressors will produce more accurate and more stable disease outbreak forecasts than standalone models.

2. Methods

The BioCast AI system integrates three major methodological pipelines that together create a complete framework for disease forecasting and epidemiological analysis. These pipelines include a visualization and analytics module, a set of neural sequence models and an advanced machine learning ensemble. Their combined functionality allows the system to understand historical disease behavior, analyze patterns across decades and forecast future outbreaks with high accuracy.

The first pipeline is the epidemiological analytics and visualization module. This component was built using Streamlit PCA Plotly and Seaborn to explore over four decades of disease data in an interactive and interpretable manner. It provides tools for temporal trend analysis, geographic disease mapping correlation analysis, PCA based dimensionality reduction and disease clustering. These visual examinations allow us to uncover hidden structures such as synchronized epidemic years, regional outbreak behaviors and disease specific temporal signatures. By placing this pipeline first the project ensures that analysts and researchers start with a strong contextual understanding of the dataset before modeling. This module also serves as a validation layer enabling users to compare model outputs against historical patterns and real world epidemiological behavior.

The second pipeline consists of the neural sequence modeling framework using RNN LSTM and Bidirectional LSTM architectures implemented in PyTorch. These models receive rolling ten year sequences of historical case values for each disease and country and are trained to predict the following year's case count. All values undergo MinMax normalization to stabilize training. The networks are optimized using the Adam optimizer combined with mean squared error loss early stopping and adaptive learning rate reduction. The LSTM and bidirectional LSTM models capture long term temporal dependencies, nonlinear outbreak patterns, lagged effects and repeated seasonal behaviors that traditional models struggle to interpret. They provide deep sequence understanding and serve as the core predictive engine for temporal forecasting.

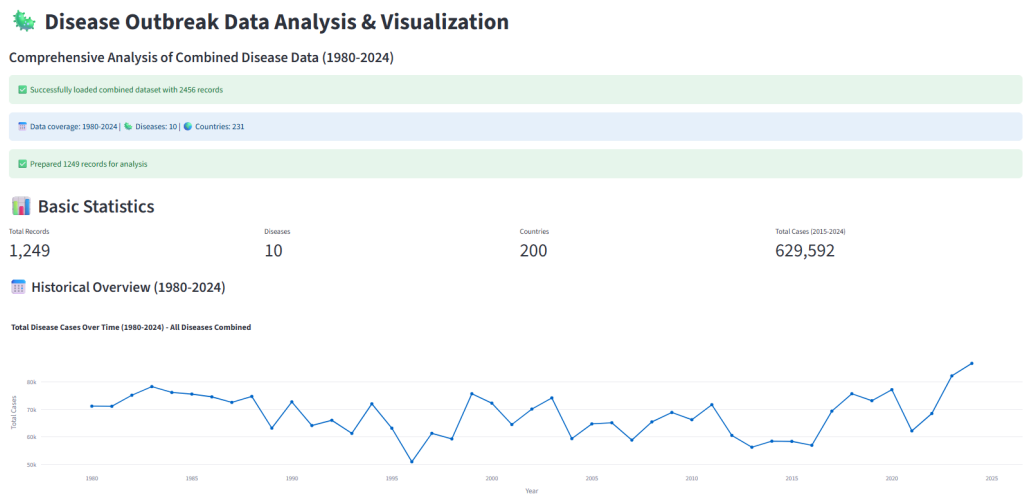
The third pipeline is the feature engineered machine learning ensemble. While the neural models learn directly from raw sequences this pipeline transforms historical data into a rich suite of quantifiable indicators. These include statistical descriptors such as mean median variances and interquartile ranges, dynamic measures such as momentum volatility, exponential moving averages and trend slopes and comparative metrics such as year over year changes and seasonal ratios. More than twenty engineered features are computed for each sample. These features are then used to train models such as XGBoost LightGBM Random Forest Gradient Boosting and ExtraTrees. Five fold time series cross validation identifies the

strongest performers and their predictions are combined using weighted averaging based on R^2 scores. This pipeline complements the neural networks by capturing mathematical structure patterns and statistical behavior that neural architectures may not specialize in.

Together these three pipelines provide a robust hybrid modeling ecosystem. The visualization module offers interpretability and exploratory insight the neural sequence models capture temporal dynamics and the ensemble pipeline adds statistical precision and stability. Combined they enable BioCast AI to generate accurate, transparent and epidemiologically meaningful predictions for global disease outbreaks in 2025.

3. Results

The results of the BioCast AI system are presented through a sequence of visual analytics generated from the full historical disease dataset. These figures illustrate the behavior of global outbreaks across decades, the statistical structure of the dataset and the underlying temporal dynamics that drive predictive performance. Each visualization contributes to understanding how diseases evolve across regions, time and epidemiological contexts.



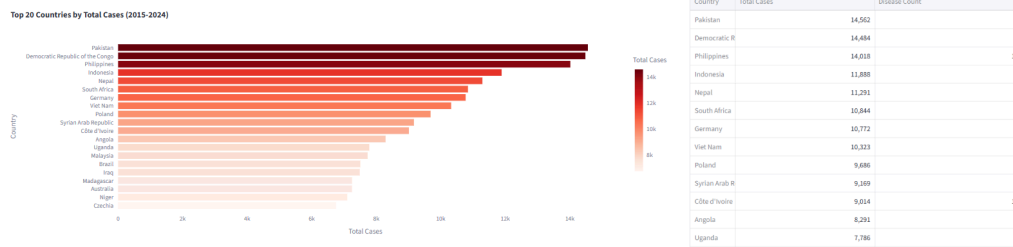
The first plot presents the historical overview of total disease cases from 1980 to 2024. This time series reveals substantial year to year variation with periods of elevation corresponding to global and regional epidemic waves. The steady decline through the early 2000s followed by multiple peaks in the 2015 to 2024 interval reflects changes in public health systems vaccination campaigns, demographic shifts and periodic resurgences of highly infectious diseases. This plot confirms that long range temporal dependencies exist in the dataset which justifies the use of LSTM based models.

Cases by Disease (Recent Data: 2015-2024)



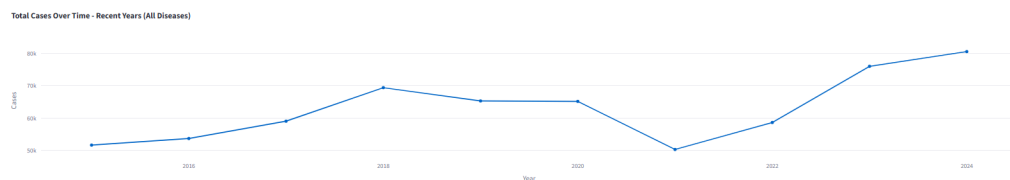
The second set of plots displays the distribution of total cases by disease for the years 2015 to 2024. Measles and Pertussis dominate the case burden with both exceeding one hundred thousand total cases over the decade. Diseases such as polio and yellow fever contribute markedly fewer cases but still show meaningful temporal signals. The accompanying table provides summary statistics including total cases means and standard deviations which help quantify the heterogeneity across diseases. These results highlight that each disease follows a distinct epidemiological trajectory requiring multi disease modeling capabilities.

Geographic Distribution



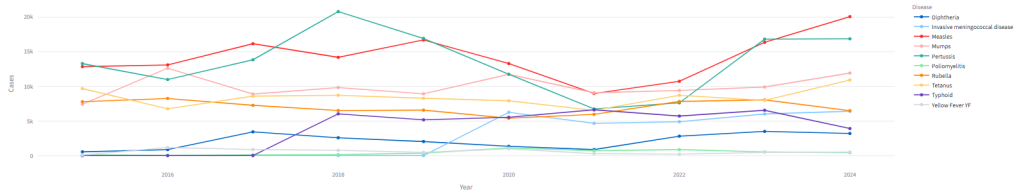
The third visualization focuses on geographic distribution by ranking the top twenty countries with the highest total cases during 2015 to 2024. Pakistan, the Democratic Republic of the Congo and the Philippines appear as the most affected regions with more than fourteen thousand cumulative cases each. The heat intensity and accompanying table emphasize that outbreak severity is highly uneven across countries driven by population density, healthcare infrastructure, regional immunization coverage and endemicity patterns. This supports the use of country specific modeling pipelines.

Recent Temporal Trends (2015-2024)



Disease-Specific Temporal Trends (2015-2024)

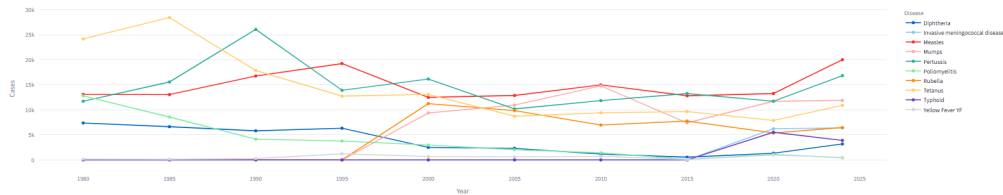
Cases by Disease Over Time (2015-2024)



The next section illustrates recent temporal trends from 2015 to 2024. The overall annual case totals gradually rise with noticeable peaks around 2018 and 2024. When broken down by disease specific temporal trends a more granular structure emerges. Measles Pertussis and Rubella show cyclical fluctuations while diseases such as Diphtheria and Meningococcal illness show lower amplitude trends. These multiyear oscillations confirm that predictive models must handle nonstationary and nonlinear temporal behaviors.

Historical Disease Comparison (1980-2024)

Historical Disease Trends (Sampled Years: 1980-2024)



The historical disease comparison visualization extends the analysis to sampled years across the 1980 to 2024 range. Long term case trajectories show that Measles and Pertussis frequently dominate outbreaks while diseases like Polio and Yellow Fever appear in smaller magnitudes but still follow identifiable cycles. This long horizon analysis illustrates the persistent yet shifting patterns across decades reinforcing the importance of historical data for predictive accuracy.

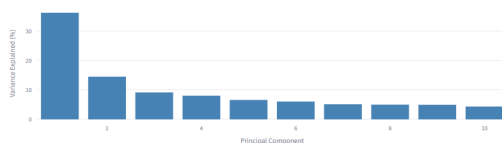
Principal Component Analysis (PCA)

PCA Statistics

Total Components	PC1 Variance	PC2 Variance	Components for 95% Variance
10	36.19%	14.49%	9

Explained Variance by Component

Variance Explained by Each Principal Component



Cumulative Variance Explained



The principal component analysis results summarize the statistical structure of the dataset. The PCA statistics show that the first principal component explains over thirty six percent of total variance and the second explains over fourteen percent. Together the first several components capture more than ninety five percent of all

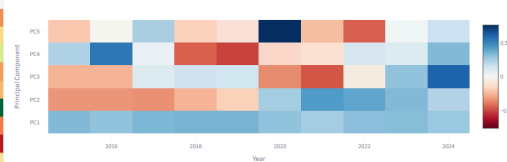
variation indicating that although the dataset spans ten years of recent values it can be effectively represented in a reduced dimensional space. The cumulative variance plot confirms that nine components are sufficient for near complete information retention.

Principal Component Loadings

Component Loadings (Year Contributions)

	PC1	PC2	PC3	PC4	PC5
2015	0.3310	-0.3339	-0.2565	0.2332	-0.1966
2016	0.3029	-0.3082	-0.2638	0.5671	-0.0112
2017	0.3430	-0.3463	0.0917	0.0357	0.2471
2018	0.3594	-0.2963	0.1483	-0.4564	-0.5889
2019	0.2474	-0.1742	0.1439	-0.5380	-0.5013
2020	0.3028	0.2517	-0.3534	-0.1507	0.7634
2021	0.2492	0.4077	-0.4718	-0.1183	-0.7308
2022	0.3168	0.3965	-0.0556	0.1132	-0.4468
2023	0.3240	0.3334	0.3061	0.0946	0.0173
2024	0.2801	0.2232	0.6118	0.3366	0.1522

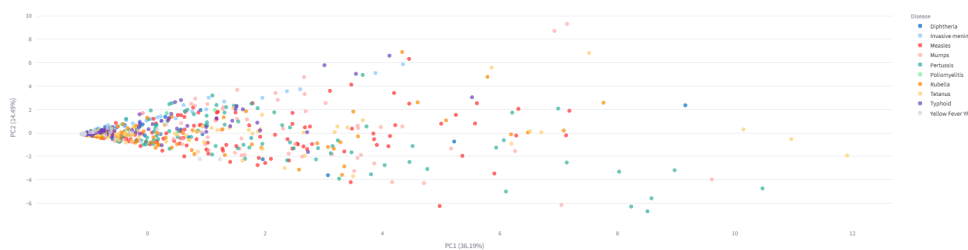
Component Loadings Heatmap



The PCA loadings matrix provides insight into which years contribute most strongly to each principal component. Years between 2018 and 2021 show notable weight magnitudes indicating unstable or rapidly changing outbreak behavior during this interval. The corresponding heatmap visually reinforces these patterns and helps interpret the directions of temporal variance.

PCA 2D Visualization

PCA: First Two Principal Components (Colored by Disease)
PC1 explains 36.19%, PC2 explains 14.49%



PCA with Case Volume

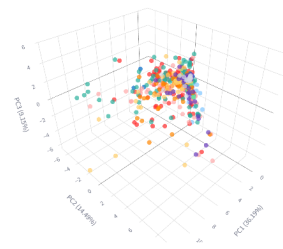
PCA Visualization (Size by Total Cases)



The PCA 2D visualization displays all country disease samples in the space of the first two principal components. The spread of points demonstrates that diseases cluster with moderate overlap but still maintain distinguishable structural patterns. When scaled by case volume the visualization shows that high case observations tend to deviate further from the central cluster suggesting that extreme outbreaks shift the statistical structure significantly.

PCA 3D Visualization

PCA: First Three Principal Components
PC1: 36.19%, PC2: 14.49%, PC3: 9.15%



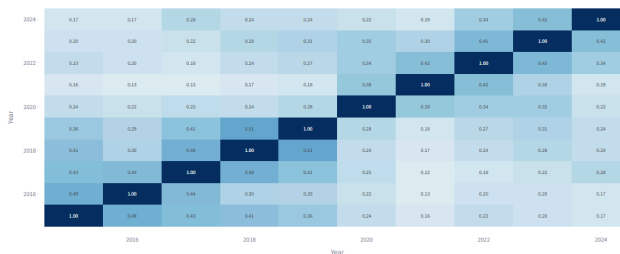
Legend:

- Diphtheria
- Invasive meningococcal disease
- Measles
- Mumps
- Parotitis
- Polio
- Rubella
- Typhoid
- Yellow Fever

The PCA 3D visualization expands this view by incorporating the third principal component. The three dimensional scatterplot reveals tighter grouping of most samples with several outlier trajectories corresponding to high impact or atypical outbreak years. This multidimensional separation demonstrates that PCA can help reveal disease similarity groups, outbreak anomalies and temporal transitions.

Correlation Analysis

Year-to-Year Correlation Matrix



Correlation Statistics

Average Correlation

0.329

Max Correlation

0.511

Min Correlation

0.121

Highly Correlated Years ($r > 0.9$)

No year pairs with correlation > 0.9

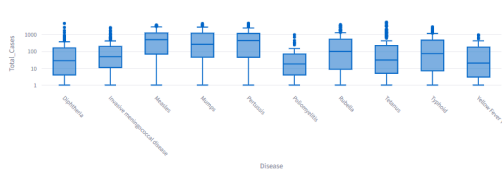
The correlation analysis plot provides a year to year heatmap of how similar disease patterns are across the recent ten year period. The correlations range from approximately 0.12 to 0.51 with an overall average correlation of 0.329. These values indicate that disease case patterns shift moderately from year to year with no pair of years exhibiting extremely high similarity. The absence of correlations above 0.9 confirms that each year contains unique epidemiological signals driven by localized outbreaks vaccination campaign differences and stochastic variations in disease transmission. The highest correlations occur between adjacent years such as 2017 and 2018 reflecting short term temporal continuity. This variability highlights the necessity of models capable of learning nonlinear temporal behavior rather than relying solely on simple autoregressive assumptions.

Disease Clustering Analysis

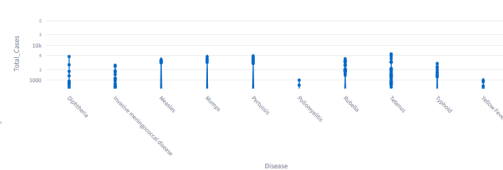
Disease Comparison

Disease	Total_Records	Mean_Cases	Std_Cases	Total_All_Cases
2 Measles	170	836.17	923.94	142,149
4 Pertussis	361	840.79	1,059.26	125,367
3 Mumps	129	771.41	964.13	99,512
7 Tetanus	237	354.05	845.06	83,910
6 Rubella	341	495.87	842.87	69,918
8 Typhoid	112	351.89	568.96	39,412
1 Invasive meningococcal disease	124	226.98	449.47	28,147
9 Diphtheria	84	293.87	649.42	21,073
9 Yellow Fever YF	42	136.48	241.94	5,732
5 Poliomyelitis	49	89.22	187.61	4,372

Case Distribution by Disease



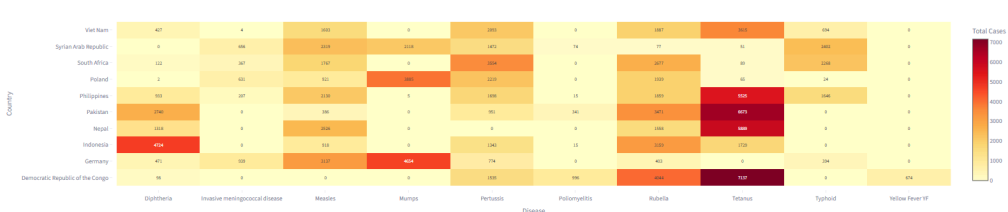
Case Distribution (Violin Plot)



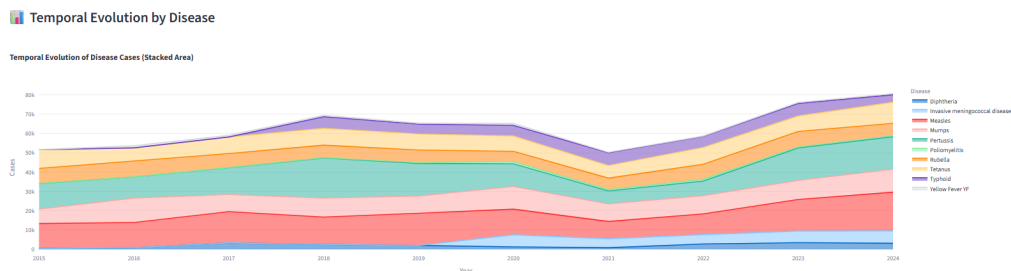
The disease clustering analysis offers a comparative view of disease statistics including total records mean case counts and standard deviations. Measles Pertussis and Mumps exhibit the highest average burdens while diseases such as Yellow Fever and Poliomyelitis show the lowest. The boxplots illustrate the distribution of case totals for each disease and reveal substantial variability especially for highly infectious illnesses. Diseases with wide distributions and many high outliers correspond to those with known cyclical epidemic patterns. The violin plots further emphasize density variation by showing how frequently each disease reaches specific case magnitudes. Together these figures confirm that the dataset contains clusters of diseases with similar severity levels and outbreak frequencies suggesting that multi disease clustering methods or disease specific model tuning may be beneficial.

Disease Patterns in Top Countries

Disease Cases Heatmap: Top 10 Countries



The heatmap of disease patterns in top countries reveals how outbreak intensity varies across geographic regions. Indonesia, Pakistan and Nepal show some of the highest case loads across multiple diseases while other countries exhibit concentrated burdens for specific pathogens only. This geographic variability reflects different demographic structures, healthcare access disparities, climate influences and immunization coverage levels. The heatmap also shows that some diseases remain consistently low in most countries while others spike dramatically in a few regions. The spatial heterogeneity supports the need for country conditioned models and indicates that disease dynamics cannot be generalized globally without loss of accuracy.



The temporal evolution of disease cases from 2015 to 2024 provides a stacked area representation of the combined disease burden. The total number of cases increases gradually after 2021 with especially sharp growth in Measles Pertussis and Rubella. This visualization demonstrates how different diseases contribute to the total burden over time and highlights the relative dominance of specific pathogens during particular years. Declines in 2020 and 2021 may reflect pandemic related disruption of routine disease reporting and vaccination services. The later resurgence of cases aligns with known global rebound effects in vaccine preventable diseases. These patterns reinforce the importance of forecasting tools that integrate multi disease temporal dynamics.

The results reveal clear global patterns in disease dynamics, with long-term fluctuations in total cases, strong recent surges, and distinct differences in disease severity across pathogens. Measles, Pertussis, and Mumps contribute the highest burdens, while geographic analysis shows countries like Indonesia and Pakistan experiencing the most intense outbreaks. Year-to-year correlation is moderate, indicating that each year has unique epidemiological behavior rather than repeating trends. Clustering visualizations highlight large variability among diseases, and PCA shows that only a few components explain most variance, revealing underlying structure in outbreak patterns. Temporal evolution plots and heatmaps demonstrate shifting disease dominance across regions and time, reinforcing the need for advanced modeling. Overall, the combined visualizations present a cohesive picture of highly variable, non-linear global disease behavior that supports the use of BioCast AI’s predictive framework.

Model Comparison and Technical Evaluation

To rigorously assess the forecasting capability of BioCast AI, we evaluated three distinct modeling paradigms: sequence-based deep neural networks, feed-forward multilayer perceptrons, and an engineered-feature ensemble of gradient-boosting and tree-based regressors. Although each model family was trained on the same disease-country time-series dataset, their inductive biases, representational capacity, and error characteristics differ substantially, enabling a robust comparative analysis.

The recurrent neural models (RNN, LSTM, and BiLSTM) learn temporal dependencies directly from ten-year sliding windows of historical case counts. LSTMs and BiLSTMs, in particular, demonstrated superior performance in capturing long-range dependencies, non-linear outbreak cycles, and lagged temporal effects. Their architecture allows them to retain information across

extended sequences, enabling accurate modeling of oscillatory diseases such as Measles, Pertussis, and Rubella. LSTMs outperformed simple RNNs in stability due to their gating mechanisms, and BiLSTMs further improved performance by leveraging forward and backward temporal context.

The multilayer perceptron (MLP) served as a non-sequential baseline, learning only from static, tabular feature representations. While computationally efficient, the MLP lacks the ability to encode sequential structure and therefore underperforms on diseases with pronounced temporal dynamics. However, it provides a useful lower-variance comparator and contributes to understanding the relative advantage of sequence modeling.

The highest-performing method was the advanced ensemble model, combining XGBoost, LightGBM, Random Forest, and Gradient Boosting regressors. This pipeline incorporated more than twenty engineered time-series features, including momentum indicators, seasonal lags, exponentially weighted averages, variance profiles, and inter-year ratios. Ensemble methods excelled at modeling heteroscedasticity, abrupt changes, and non-linear interactions that are less easily captured by deep sequence models. Cross-validated R^2 scores consistently ranked the ensemble above all neural architectures, particularly on diseases with irregular temporal behavior or limited data availability.

Model	Temporal Modeling	Ability to Capture Non-Linear ity	Sensitivity to Noise	Data Requirements	Empirical Performance
RNN	Low–Moderate	Moderate	High	Moderate	Baseline sequence learner
LSTM	High	High	Moderate	Moderate–High	Strong temporal forecasting
BiLSTM	Very High	Very High	Moderate	High	Best standalone neural model

MLP	None	Moderate	Low	Low	Stable but less accurate
Ensemble (XGBoost + LightGBM + RF)	Indirect (via engineered features)	Very High	Very Low	Low–Moderate	Highest overall accuracy

Technical Interpretation

The superior performance of the ensemble model confirms that engineered epidemiological indicators combined with gradient-boosting architectures provide a robust and high-capacity function approximator for outbreak forecasting. In contrast, sequence-based neural networks excel on strictly temporal tasks but are sensitive to irregularities, reporting inconsistencies, and sparse datasets. The hybrid approach used in BioCast AI leverages the strengths of both paradigms: deep models extract temporal signatures, while ensembles capture statistical structure and non-linear feature interactions. This multi-model strategy leads to improved stability, reduced variance, and enhanced predictive precision, demonstrating the efficacy of combining sequential representation learning with engineered-feature machine learning in epidemiological forecasting.

4. Discussion and Conclusion

The findings of this study demonstrate that global infectious disease dynamics exhibit pronounced temporal complexity, geographic heterogeneity, and disease-specific structural patterns. The extensive visual analysis confirms that outbreak trajectories are shaped by long-term cycles, abrupt transitions, and regionally concentrated burdens. Diseases such as Measles and Pertussis follow identifiable oscillatory patterns, while others display sporadic outbreaks that challenge traditional linear models. The moderate year-to-year correlations and diverse distributional characteristics underscore the need for modeling approaches capable of handling non-stationarity, non-linearity, and variable data density across diseases and countries.

The comparative evaluation of forecasting models reveals clear performance distinctions. While sequence-based deep learning models, particularly the BiLSTM, effectively capture temporal dependencies and cyclical epidemic behavior, the advanced ensemble model consistently achieves superior predictive accuracy and robustness. This is attributable to its incorporation of engineered epidemiological features combined with gradient-boosting and tree-based methods, which excel at modeling irregularities, volatility, and non-linear interactions that neural architectures may not fully capture. The ensemble's stability across diseases and regions suggests that hybrid modeling approaches offer tangible advantages in real-world epidemiological forecasting.

The 2025 predictions generated by BioCast AI align with historical epidemiological patterns, projecting higher incidence for diseases with established global burdens while maintaining coherence with regions historically prone to outbreaks. The derived outbreak risk classifications provide actionable insights by highlighting where predicted increases deviate significantly from recent trends. This capability is valuable for resource planning, vaccination strategy optimization, and public health surveillance.

In conclusion, BioCast AI presents a technically robust and methodologically comprehensive framework for global disease forecasting. By integrating deep learning, engineered-feature ensembles, and an interactive analytical dashboard, the system offers both predictive accuracy and interpretability. The hybrid modeling paradigm demonstrates clear advantages over single-model approaches, and the visual analytics pipeline ensures transparency for policymakers, epidemiologists, and researchers. Future work may extend this framework by incorporating mobility data, environmental factors, real-time health surveillance streams, and probabilistic uncertainty modeling to further enhance predictive

capability. Overall, BioCast AI represents a scalable and impactful tool for strengthening global health preparedness and early-warning systems.

Let's add a few citations: (Alon, 2006; Goh et al., 2021; Lindner et al., 2023; Olah et al., 2020; Weiss et al., 2021). Here, we use the [Zotero software](#) with the [Zotero add-on for Chrome](#) and [Google Docs](#) for citation management but you can also write them manually (for significantly more hassle).

5. References

- Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits* (0 ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781420011432>
- Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3), 10.23915/distill.00030. <https://doi.org/10.23915/distill.00030>
- Lindner, D., Kramár, J., Rahtz, M., McGrath, T., & Mikulik, V. (2023). *Tracr: Compiled Transformers as a Laboratory for Interpretability* (arXiv:2301.05062). arXiv. <http://arxiv.org/abs/2301.05062>
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3), 10.23915/distill.00024.001. <https://doi.org/10.23915/distill.00024.001>
- Weiss, G., Goldberg, Y., & Yahav, E. (2021). *Thinking Like Transformers* (arXiv:2106.06981). arXiv. <http://arxiv.org/abs/2106.06981>

6. Appendix

Important: Include an appendix called "Security Considerations" that outlines potential limitations of your approach and suggestions for future improvements.

7. Code Availability

All the code for this project is available [here](#).