Forecasting Chronic Wasting Disease with a Metapopulation-SIR Informed Graph Neural Network

Victor Henriksson*
Georgia Institute of Technology
Atlanta, GA
vhenriksson3@gatech.edu

Jack Keller
Georgia Institute of Technology
Atlanta, GA
jkeller44@gatech.edu

Joseph Yoo
Georgia Institute of Technology
Atlanta, GA
jyoo339@gatech.edu

Figure 1: An elk infected with Chronic Wasting Disease (CWD), a degenerative disease which affects members of the deer family. It is always fatal, and there is no known treatment [8].

Abstract

Chronic Wasting Disease (CWD) is a fatal, prion-based disease that affects deer populations (i.e., cervids) worldwide. As the disease spreads, an important question arises: how will it influence human behavior, particularly the hunting and consumption of cervids – especially white-tailed deer in the Midwest and Southeast U.S. – through disease-management policies? We propose a hybrid machine learning-mechanistic model to forecast the spread of CWD in these regions to better understand its potential impact on human behavior. The model consists of a graph neural network (GNN) whose loss function incorporates a Metapopulation SIR model using the physics-informed neural network (PINNs) framework. This approach captures both the spatial and temporal dynamics of disease transmission while remaining grounded in epidemiological constraints. We find improved accuracy using the proposed model against a classical Metapopulation-SIR ODE model and other neural network methods, which allows for better prediction of high infection count counties.

CCS Concepts

• **Computing methodologies** → **Network science**; **Machine learning**; **Modeling and simulation**; • **Applied computing** → **Health informatics**.

Keywords

Chronic Wasting Disease (CWD), Epidemic Forecasting, Graph Neural Networks (GNNs), Physics-Informed Neural Networks (PINNs), Metapopulation Models, SIR Models

1 Introduction

Chronic Wasting Disease (CWD) is a degenerative disease which affects cervids, which are members of the deer family including deer, elk, and moose. Reported primarily in North America, it has been reported in 36 U.S. states (see Figure 2), as well as Canada, Norway, Finland, and Sweden [4]. It is a prion disease, which is caused by misfolded proteins, and primarily attacks the nervous system. As the disease progresses, symptoms include stumbling, tremors, excessive drooling, loss of fear of humans, and eventually death. There is no known cure or treatment [3].

*All authors contributed equally to this research.

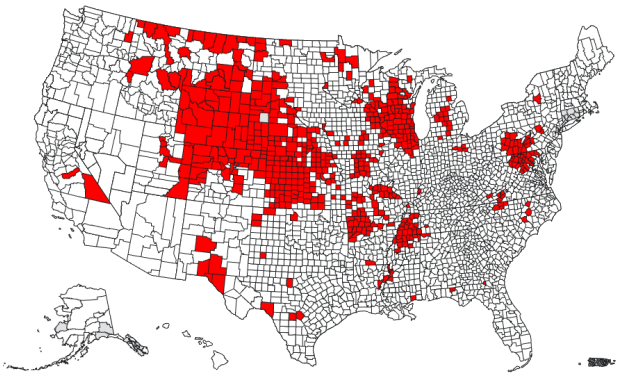


Figure 2: Counties in the United States with reported cases of CWD in wild cervids, as of April 2025. Data is based on state wildlife agencies and the United States Geological Survey [4].

CWD was first detected in the wild in 1981, and it has continued to spread since. Additionally, CWD is highly contagious – thought to spread between animals through contact with saliva or other bodily fluids, or indirectly through the environment such as in soil or drinking water, it's estimated that statewide rates of CWD in the U.S. may exceed 1 in every 10 susceptible animals. Along with the fact that CWD can linger within soil or water for years, another difficult challenge of fighting CWD is that it can take months, or even years, before infected animals begin to display symptoms [3].

Along with the obvious harm the disease causes to cervids, CWD imposes a heavy economic burden as well. In a 2002 study, it was found that the discovery of CWD infections decreased hunter participation by 5.4% in Wisconsin [6]. This is significant – not only is hunting crucial for regulating cervid populations, but it also supports industries such as hospitality and dining, especially in rural areas. Another study found that the farmed cervids industry spent $\$307,950$ in 2020 on sampling for CWD, and the US government spent $284.1 million from 2000 to 2021 in efforts to combat the spread of CWD [5].

2 Problem Definition

Given the current difficulties of containing the spread of CWD and its devastating impacts on cervid populations and dependent economies, this paper seeks to develop a state-of-the-art forecasting model for CWD to enable governments and farmers to better respond proactively and limit the spread of the disease. More formally, we seek to address the following research question: *Can the spread of Chronic Wasting Disease be accurately predicted both across and within regions over time?* To do so, we employ a graph neural network (GNN), informed by a metapopulation-SIR model tailored to CWD infection dynamics, to predict the spread of CWD throughout different regions over time. A detailed description of our modeling framework and applied techniques is provided in Section 4.

The impact of this model is twofold: it advances epidemic forecasting through a novel integration of GNN and a domain-specific metapopulation-SIR model, and it provides governments, farmers, and hunters with a practical tool to better anticipate and mitigate the spread of CWD.

3 Related Work

This proposal builds on the previous work in the field of forecasting diseases and Chronic Wasting Disease. Only recently, CWD datasets and machine learning techniques have been used to forecast the spread of the disease. A 2024 paper presents a large dataset of 31,636 entries from June 2000 - 2022 of county level infections of CWD for the north and south east of the United States. Using this dataset, the researchers tested four machine learning techniques: random forest, decisions trees, gradient boosting, and light gradient boosting. However after preparing the data for usage with the aforementioned methods, the usable dataset fell to 158 entries for 2019-2020. This issue could be addressed by using the entire dataset to train the model, then forecast infections from new data. Additionally, the paper states the recommendation that "future ML models better characterize the spread of disease across the landscape by incorporating geographical proximity," which is addressed through the usage of a graph and graph neural networks for the county level [1]. The graph networks inherently model the geographical proximity.

Additionally, similar models have already been successfully applied to other diseases in epidemic forecasting. For instance, researchers at Shanghai University and Zhejiang Laboratory in China aimed to embed epidemiological domain-knowledge regarding metapopulations into a spatio-temporal model. They developed a hybrid model called MPSTAN in which the loss function was based on a combination of the predictions of a traditional metapopulation-based SIR model and a graph neural network. In doing so, they achieved state-of-the-art performance on two real-world COVID-19 datasets. Strengths of this approach were excellent performance, a relatively small number of parameters, and the successful incorporation of epidemiological domain knowledge [7]. However, this work could be extended by experimenting with different graph constructions and SIR models designed to the specific problem domain within epidemiology, which is what we aim to do.

Another approach published in 2024 aimed to improve epidemic forecasting accuracy by simultaneously learning spatio-temporal (i.e., observed) and mechanistic (e.g., infection rates) features. Similar to MPSTAN, they developed a hybrid loss function which incorporates predictions from different methods; however, the backbone of their model differs from MPSTAN. While the model was generally very accurate and robust to different forecasting horizons and datasets, there was some performance drop for long-term forecasting horizons [10]. Additionally, the model was limited to only using the traditional SIR model, which could be extended upon.

The work done on metapopulation-based epidemiology studies has allowed for a better understanding of disease spread through communities and regions. The fitting of parameters of the SIR models and transitions between regions is addressed in the following paper via the use of machine learning techniques like graph neural

networks (GNNs) [2]. The benefits of this model, MepoGNN, is that its predictions take into account spatial and temporal movement of the disease throughout regions. The model works by predicting the parameters of infection rate and removal rate for each region and timestep of the metapopulation SIR. Additionally, the model learns to predict the mobility of individuals in the metapopulation study. Using these parameters, the model can directly use the metapopulation SIR model to predict the number of cases daily. Through an ablation study, the paper showed that the use of the metapopulation SIR model directly in the model increased the performance. However, the model does not perform well with the early stages of an epidemic. For CWD, this issue is addressed by the size and time the dataset covers. Overall, this paper demonstrates that SIR-GNN based models can perform well in metapopulation based studies.

Physics-informed neural networks (PINNs) have also been applied to epidemic forecasting by encoding differential-equation constraints directly in the loss [11]. Similarly, Di Lorenzo et al. [6] showed that combining physics-based constraints with GNNs improves interpretability and spatial inference. Together, these studies highlight the benefit of embedding mechanistic knowledge within data-driven architectures, a core idea our metapopulation-SIR GNN extends to a county-level wildlife disease context.

4 Methodology

4.1 Data Collection and Graph Construction

We utilize a comprehensive CWD surveillance dataset aggregating 31,636 county-level records from July 2000 to June 2022 across the Midwest and Southeast United States [1]. The data covers 1,438 counties across 16 states (e.g., Wisconsin, Iowa, Tennessee) and includes annual counts of positive and negative CWD tests, cervid harvests, and captive cervid facilities. To capture the spatial dynamics of disease spread, we construct a static graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each node $i \in \mathcal{V}$ represents a county. Unweighted edges $(i, j) \in \mathcal{E}$ connect geographically adjacent counties based on U.S. Census Bureau data [9], resulting in a graph with $|\mathcal{V}| = 1438$ nodes and $|\mathcal{E}| = 8064$ edges. This graph is sparse ($|\mathcal{E}| \approx 5.6|\mathcal{V}|$). Data cleaning and graph construction are implemented in `clean_cwd_data.py` and `build_network.py`.

4.1.1 Feature Representation. For each county i and season τ, we construct feature vectors to represent environmental stability and surveillance intensity.

- **Static Features ($\mathbf{x}_i^{(s)} \in \mathbb{R}^3$):** Geodesic centroid (latitude, longitude) and land area, representing fixed spatial characteristics.
- **Temporal Features ($\mathbf{x}_{i,\tau}^{(t)} \in \mathbb{R}^4$):** A dynamic vector consisting of annual counts for (1) positive CWD tests, (2) negative CWD tests, (3) total cervid harvest, and (4) number of cervid facilities.

Missing temporal entries are imputed with zeros. To address the extreme scale imbalance—where harvest counts ($\sim 10^4$) dwarf infection counts ($\sim 10^0$)—we apply a $\log(1 + x)$ transform to the infection-related channel (positive tests). Total harvest counts are normalized by the maximum observed harvest ($\approx 2.1 \times 10^4$). This ensures all inputs occupy comparable numerical ranges before entering the encoder.

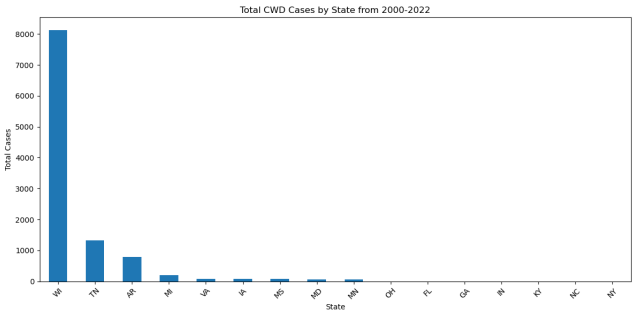


Figure 3: This figure compares the total CWD cases in each state from July 2000 - June 2022. Wisconsin has nearly 8x the number of cases than the second-most state, Tennessee.

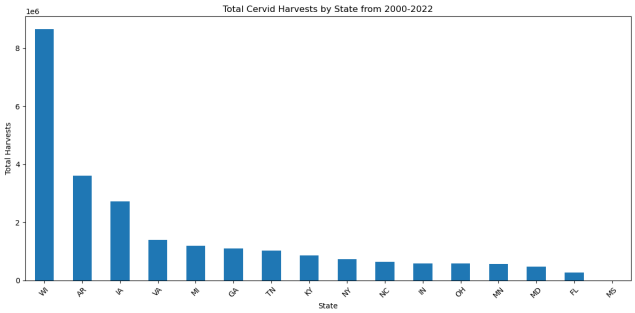


Figure 4: This figure compares the total cervid harvests in each state from July 2000 - June 2022. Both Wisconsin and Iowa are in the top three.

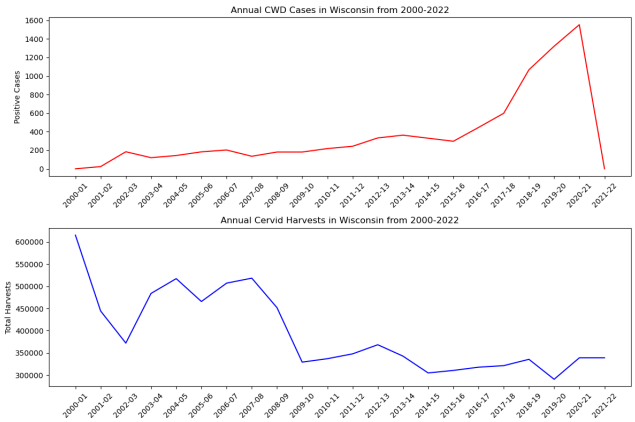


Figure 5: Annual CWD cases and cervid harvests in Wisconsin from July 2000 - June 2022. Notably, there has been a large spike in CWD cases in recent seasons.

4.1.2 Target Definition and Masking. Our primary forecasting target is the *infection prevalence* (fraction of the tested population that

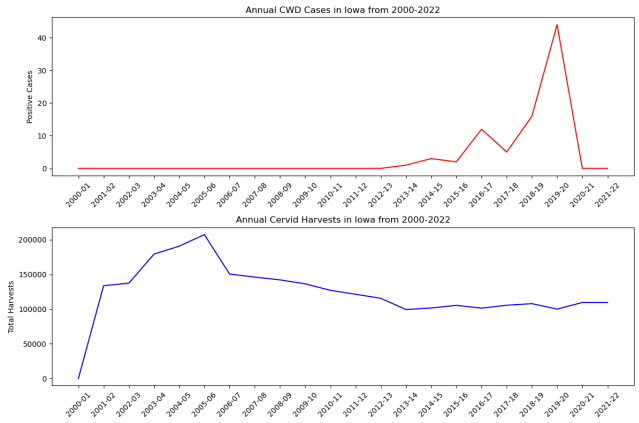


Figure 6: Annual CWD cases and cervid harvests in Iowa from July 2000 - June 2022. Notably, there has been a large spike in CWD cases in recent seasons.

is positive). We define the ground-truth target $y_{i,\tau} \in [0,1]$ as:

$$y_{i,\tau} = \frac{\text{pos}_{i,\tau}}{\text{pos}_{i,\tau} + \text{neg}_{i,\tau} + \varepsilon} \tag{1}$$

Surveillance intensity is highly non-uniform, and many county-years have no tests at all. To restrict supervision to meaningful observations, we construct a binary label mask

$$m_{i,\tau} = \mathbb{I}\big[\text{pos}_{i,\tau} + \text{neg}_{i,\tau} > 0\big], \tag{2}$$

where $m_{i,\tau} = 1$ if county i has at least one CWD test in season τ and $m_{i,\tau} = 0$ otherwise. All infection- and physics-related losses in Section 4.2 are computed only over county-years with $m_{i,\tau} = 1$, preventing the model from being biased by structurally missing data.

4.1.3 Experimental Setup. We employ a temporal hold-out evaluation strategy. A surveillance "season" is defined as the span from July 1st to June 30th (e.g., the 2000–2001 season). The model is trained on historical data from the 2000–2001 season through the 2019–2020 season across all 16 states. Evaluation is performed on the held-out 2020–2021 season. We exclude the 2021–2022 season entirely due to incomplete reporting (missing data) in the source dataset. Figures 5 and 6 summarize temporal trends in reported infections and harvest in Wisconsin and Iowa, illustrating both the emergence of sustained CWD clusters and the variability in surveillance intensity over time.

4.2 Model

We develop a hybrid graph neural network that couples a spatio-temporal GNN encoder with a county-level SIR process. The model operates on the static county adjacency graph described in Section 4.1, where nodes represent counties with local cervid populations and edges represent geographic adjacency. Each node carries static covariates and yearly CWD testing and harvest activity. The goal is to forecast infection prevalence (infection fraction) in each county for a target season while remaining consistent with SIR dynamics.

4.2.1 Input Preprocessing. Each node carries static features (latitude, longitude, area), temporal features (CWD tests, harvest, facilities), and an infection-fraction target $y_{i,\tau}$ computed from positive and negative tests (see Section 4.1). We apply a $\log(1+x)$ transformation specifically to the positive test count channel of the input temporal tensor $\mathbf{x}^{(t)}$, while keeping the infection-fraction targets $\mathbf{y}$ in their original probability space.

4.2.2 Spatio-temporal encoder. The core of the model is a spatio-temporal encoder that first mixes information across neighboring counties and then across years.

- **Spatial encoder (GATv2).** For each year τ, we concatenate static and temporal inputs and project to a hidden dimension:

$$\mathbf{h}_{i,\tau}^{(0)} = f_{\text{lin}}\big([\mathbf{x}_i^{(s)} \,;\, \tilde{\mathbf{x}}_{i,\tau}^{(t)}]\big) \in \mathbb{R}^d.$$

  These node embeddings are passed through a stack of GATv2 layers over the county graph, producing spatially enriched embeddings $\mathbf{h}_{i,\tau}^{\text{GAT}}$ that aggregate information from adjacent counties via learned attention weights.
- **Temporal processor (GRU).** For each county i, the sequence of GAT-encoded embeddings $\{\mathbf{h}_{i,1}^{\text{GAT}}, \ldots, \mathbf{h}_{i,T}^{\text{GAT}}\}$ is treated as a time series and passed through a Gated Recurrent Unit (GRU). Using nodes as the batch dimension and years as the sequence dimension, the GRU summarizes the historical trajectory of each county into a final hidden state

$$\mathbf{h}_i^{\text{enc}} = \text{GRU}\big(\mathbf{h}_{i,1:T}^{\text{GAT}}\big) \in \mathbb{R}^{d_h},$$

  which encodes both spatial context and temporal trends up to the prediction year t.

4.2.3 Multi-head decoder. To capture different epidemiological quantities while avoiding mode collapse, the final encoder state $\mathbf{h}_i^{\text{enc}}$ is fed into three task-specific regression heads:

$$\textbf{Infection head:} \quad \hat{z}_i = f_{\text{inf}}(\mathbf{h}_i^{\text{enc}}) \in \mathbb{R}, \tag{3}$$

$$\textbf{Harvest head:} \quad \hat{H}_i = f_{\text{harv}}(\mathbf{h}_i^{\text{enc}}) \in \mathbb{R}_+, \tag{4}$$

$$\textbf{Parameter head:} \quad [\beta_i, \gamma_i] = f_{\text{par}}(\mathbf{h}_i^{\text{enc}}) \in (0,1)^2. \tag{5}$$

The infection head operates in log-space for numerical stability, and we interpret $\hat{z}_i$ as the predicted log infection fraction, i.e., $\hat{z}_i \approx \log(y_{i,t} + \varepsilon)$. The harvest head predicts total harvest (ReLU constrained), and the parameter head outputs county-level transmission and removal rates (Sigmoid constrained).

4.2.4 Physics-informed hybrid loss. Training uses a composite objective that balances data fidelity with epidemiological structure. For a given prediction year t and its immediate predecessor $t-1$, the total loss is

$$\mathcal{L} = \lambda_{\text{inf}}\mathcal{L}_{\text{inf}} + \lambda_{\text{harv}}\mathcal{L}_{\text{harv}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{6}$$

We utilize the binary data availability mask $m_{i,t}$ (see Section 4.1) so that infection- and physics-related terms are computed only for county-years with valid testing data, while harvest loss is applied over all counties.

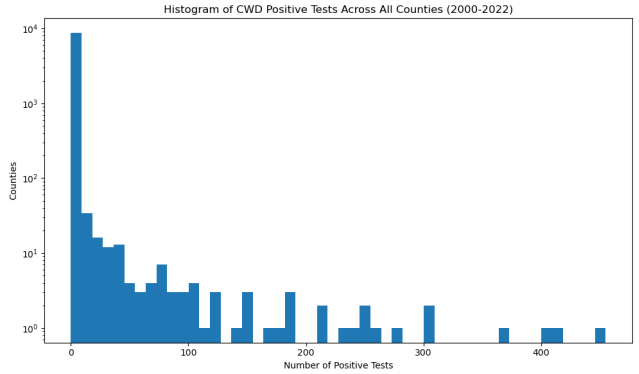


Figure 7: Histogram of positive CWD tests (log-scale) from 2000-2022, across all counties. Note the extremely high number of zeros.

Weighted infection loss $\mathcal{L}_{inf}$. Because most county-years have zero detected infections (see Figure 7), an unweighted MSE encourages a trivial all-zero solution. We instead optimize a weighted MSE in log-space:

$$\mathcal{L}_{\text{inf}} = \frac{1}{\sum_i m_{i,t} w_i} \sum_i m_{i,t}\, w_i \left\| \hat{z}_i - \log\left(y_{i,t} + \varepsilon\right) \right\|^2, \tag{7}$$

where $w_i = w_{\text{pos}}$ if $y_{i,t} > 0$ and $w_i = w_{\text{zero}}$ otherwise, with $w_{\text{pos}} \gg w_{\text{zero}}$ (e.g., $w_{\text{pos}} = 50$, $w_{\text{zero}} = 1$). This forces the model to fit rare positive-infection counties without being dominated by zeros.

Harvest loss $\mathcal{L}_{harv}$. We use a standard MSE between predicted and observed harvest counts $H_{i,t}$:

$$\mathcal{L}_{\text{harv}} = \frac{1}{N} \sum_i \left\| \hat{H}_i - H_{i,t} \right\|^2, \tag{8}$$

anchoring the model's understanding of sampling effort and population scale.

Physics residual $\mathcal{L}_{phys}$. To tie the learned parameters (β_i, γ_i) to Metapopulation-SIR dynamics, we enforce a one-step Metapopulation-SIR consistency condition. Using the ground-truth infection fraction at the previous year, $I_{i,t-1} = y_{i,t-1}$ (as defined in Section 4.1), and approximating susceptibility as $S_{i,t-1} \approx 1 - I_{i,t-1}$, we apply the following update at time t:

$$S_{i,t} = S_i^{\text{eff}} - \beta S_i^{\text{eff}} I_i^{\text{eff}}, \quad I_{i,t} = I_i^{\text{eff}} + \beta S_i^{\text{eff}} I_i^{\text{eff}} - \gamma I_i^{\text{eff}}, \quad R_{i,t} = R_i^{\text{eff}} + \gamma I_i^{\text{eff}}$$

The effective populations incorporate inter-county movement and are defined as follows:

$$S_i^{\text{eff}} = S_{i,t-1} + \Delta S_i^{\text{move}}, \quad I_i^{\text{eff}} = I_{i,t-1} + \Delta I_i^{\text{move}}, \quad R_i^{\text{eff}} = R_{i,t-1} + \Delta R_i^{\text{move}}$$

Lastly, the movement term ΔX_i^{move} for any compartment $X \in \{S, I, R\}$ is:

$$\Delta X_i^{\text{move}} = \sum_j A_{ij} X_{j,t-1} \sigma_j - X_{i,t-1} \sigma_i$$

where $A_{ij} = 1$ if nodes i, j are adjacent and $\sigma_i = \frac{1}{D_i} \cdot s$. $D_i =$ out-degree of node i, and s is a small scaling factor, which is designed to limit movement to a reasonable rate. Note that since CWD is always fatal, $R_i(t)$ represents the deceased population of a given county over time.

The physics loss compares this to the observed infection fraction:

$$\mathcal{L}_{\text{phys}} = \frac{1}{\sum_i m_{i,t}} \sum_i m_{i,t} \left\| I_{i,t} - y_{i,t} \right\|^2. \tag{9}$$

This encourages (β_i, γ_i) to explain the observed year-to-year change in prevalence.

Regularization $\mathcal{L}_{reg}$. Finally, we impose an L_2 penalty on the epidemiological parameters to control magnitude and improve interpretability:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_i \left(\beta_i^2 + \gamma_i^2 \right). \tag{10}$$

The full architecture and data flow are summarized in Figure 8, which highlights the separation between log-scaled inputs and physical targets, the spatio-temporal encoder, and the three-head decoder trained under the hybrid physics-informed objective.

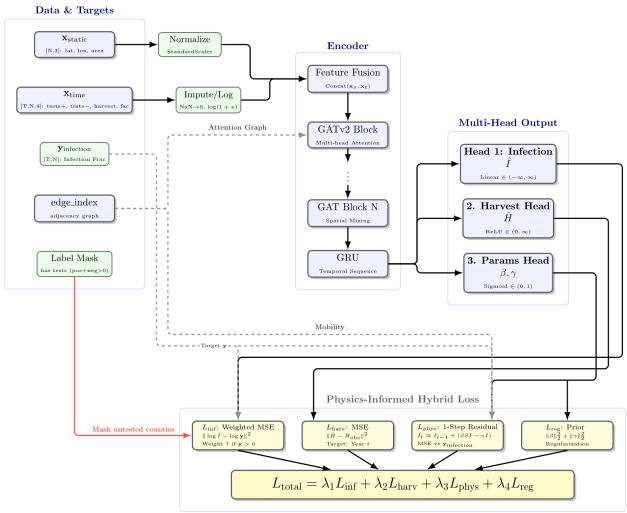


Figure 8: Overview of the Metapopulation-SIR informed Graph Neural Network (GNN) architecture.

4.3 Evaluation

We assess model performance through a structured ablation study comparing different loss configurations on a temporal hold-out set. All models are trained on sequences ending in the 2019–2020 season and evaluated on the 2020–2021 season across all 16 states.

4.3.1 Experimental Conditions. To quantify the impact of physics-informed regularization and data supervision, we compare three primary configurations:

(1) **Baseline (Harvest-Only):** Trained with $\lambda_{\text{inf}} = \lambda_{\text{phys}} = 0$, optimizing only for harvest prediction. This serves as a control to measure if infection dynamics can be learned solely from shared encoder features without direct supervision.
(2) **Data-Driven:** Trained with $\lambda_{\text{inf}} > 0$ and $\lambda_{\text{harv}} > 0$, but without physics constraints ($\lambda_{\text{phys}} = 0$). This tests the model's ability to learn infection patterns purely from sparse surveillance data.
(3) **Physics-Informed (Full):** The complete model trained with all loss terms active, enforcing both data fidelity and SIR consistency via the one-step residual.

We additionally benchmark these against standard autoregressive (AR1) and Random Forest baselines trained on the same feature set.

4.3.2 Quantitative Metrics. Our primary metric is the Mean Absolute Error (MAE) between the predicted infection prevalence $\hat{I}_{i,\tau}$ and the ground truth $y_{i,\tau}$ for the held-out target season τ^*:

$$\text{MAE} = \frac{1}{\sum_i m_{i,\tau^*}} \sum_i m_{i,\tau^*} |\hat{I}_{i,\tau^*} - y_{i,\tau^*}|. \tag{11}$$

Metrics are computed exclusively on county-years with valid surveillance data ($m_{i,\tau^*} = 1$). To assess regional generalization, we report MAE separately for the high-surveillance states (Wisconsin and Iowa) versus the remaining 14 states.

4.3.3 Epidemiological Consistency and Diagnostics. Beyond error metrics, we evaluate the biological plausibility of the learned model. We inspect the distributions of the inferred parameters β_i (transmission) and γ_i (removal) to ensure they lie within realistic ranges. We also generate scatter plots of predicted versus observed prevalence to detect mode collapse (e.g., predicting mean prevalence everywhere). Finally, we visualize spatial error maps to verify that the physics-informed model correctly identifies known infection clusters rather than smoothing them out.

5 Experiments

We propose the following experiments to evaluate the performance of the proposed network against the ground truth. We compare the proposed model against a Metapopulation-SIR ODE model, first-order auto regressive model (AR1), random forest, a graph neural network trained on only infections (Base), a graph neural network trained on infections and harvest data, a graph neural network trained on infections, harvest data, and with a Metapopulation-SIR loss term, and a graph neural network trained on infections with a Metapopulation-SIR loss term. The metric of Mean Absolute Error (MAE) is used to compare the predicted infections counts per county against the ground truth infections count of the county.

6 Results

We evaluate the infection count of CWD disease across multiple states in the southeast and northeast of the United States across 20 years from 2000 to 2021. The model is trained on the years 2000-2019 and predicted on the 2020-2021 deer hunting season. In Figure 10, we show the predicted infection count per county during the prediction period of 2020 to 2021 in the log scale. Next, we show the difference between the predicted and ground truth infection counts per county in Figure 11. Additionally, we show the error metric of MAE in Table 1. Notably, the graph neural network trained on infections with a Metapopulation-SIR loss term had the lowest MAE, which is supported in Figure 9.

Model	MAE
Base (infection data)	2.93
Base w/ harvest data	2.15
Base w/ harvest data & Meta-SIR loss	2.11
Base w/ Meta-SIR loss	**0.73**
AR(1)	1.08
Metapopulation SIR	2.62
Random Forest	2.61

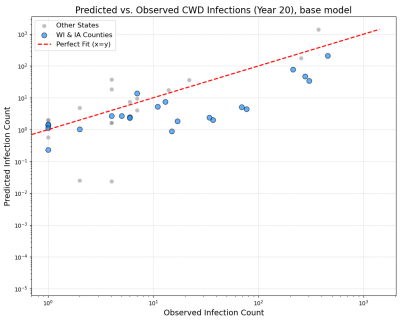
Table 1: Mean Absolute Error (MAE) scores for different models.

7 Discussion

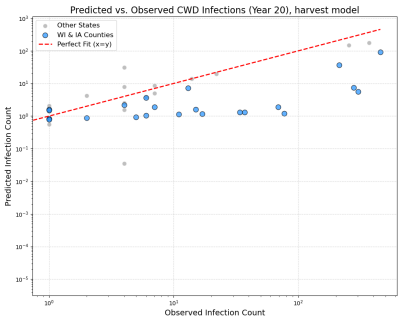
The experiments demonstrate that the base model with the Metapopulation-SIR loss has the lowest MAE against the ground truth as seen in Table 1. This model outperforms both the classical Metapopulation-SIR ODE model and other neural network methods. All models displayed in Figure 9 were able to find primary areas of infection and demonstrated some ability to model the spread. However, there is a notable issue with the neural network based models: they over-aggressively predict the spread across states, even when given a Metapopulation-SIR loss term to help regulate this. This issue is highlighted in Figure 11, as a large number of the predicted areas differ from the ground truth, as well as from the metapopulation-SIR ODE model. Furthermore, the dataset is largely filled with zero infections for a majority of the dataset (see Figure 7). This leads all the models except the Metapopulation-SIR ODE model to struggle handling the sparsity of values and the large differences of infection counts between counties. This means that the neural network models struggle to understand the spread of CWD across regions in sparse data situations. Additionally, the lack of the data from states in this dataset has led to an incomplete picture for the models to work on as seen with the absences of Illinois, Indiana, Alabama, Pennsylvania, New Jersey, Delaware, Maryland, and West Virginia in the predictions. In terms of human consumption and harvesting, the model's ability to show the regions of interest via high infection counts still allows for proper game management via culling and testing, which can protect people from exposure to CWD. However, the neural networks' problems with overly-aggressive spread hamper the ability for proper identification of movement of CWD in comparison to the Metapopulation-SIR ODE model, which shows a more realistic spread of CWD.

8 Conclusion

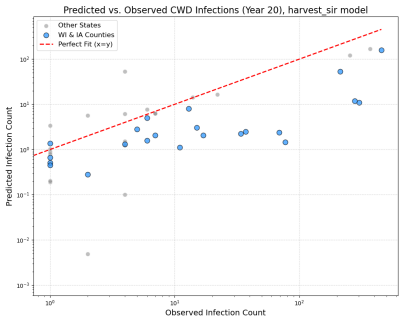
In this study, we demonstrate that a Graph Neural Network (GNN) trained on infection data with a Metapopulation-SIR loss term is able to improve the MAE of infection count of CWD across multiple states, versus a classical Metapopulation-SIR ODE model and other neural network methods. This improved model allows for a better understanding of CWD infection counts and spread, which in turn allows decision-makers to reduce the effects of CWD on both cervid



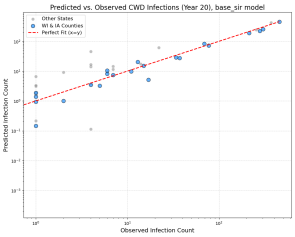
Base (infection data)



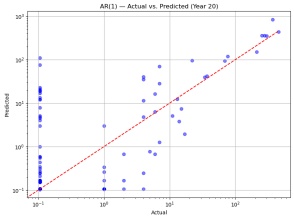
Base w/ harvest data



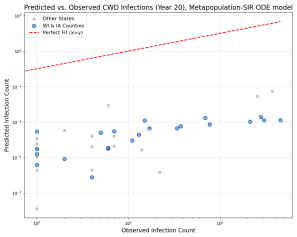
Base w/ harvest data & Meta-SIR loss



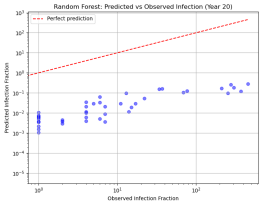
Base w/ Meta-SIR Loss



AR1

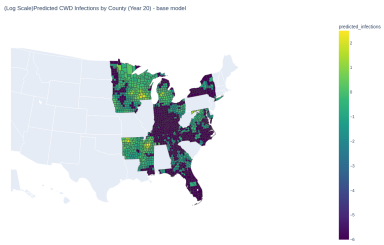


Metapopulation-SIR ODE

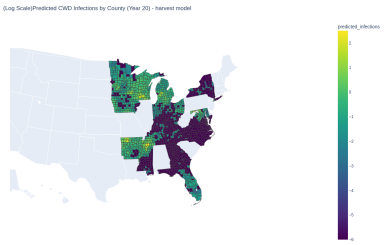


Random Forest

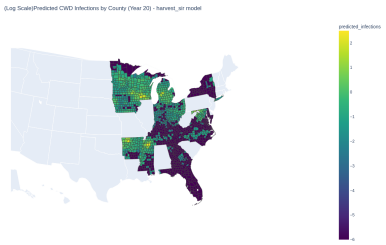
Figure 9: Predicted vs. observed infection count per county for 2020-2021 season.



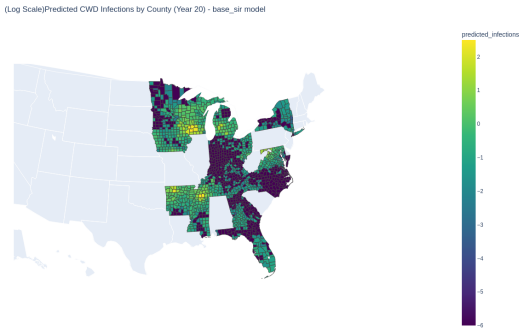
Base (infection data)



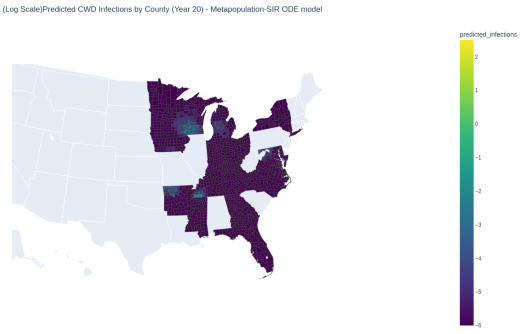
Base w/ harvest data



Base w/ harvest data & Meta-SIR loss



Base w/ Meta-SIR loss



Metapopulation-SIR ODE

Figure 10: Predicted infection count per county for 2020-2021 season (log scale).

Base (infection data) Base w/ harvest data Base w/ harvest data & Meta-SIR loss

Base w/ Meta-SIR loss Metapopulation-SIR ODE

Figure 11: Difference of infection count from predicted and ground truth for 2020-2021 season (log scale).

and human populations. However, the sparsity of the dataset limits the reliability of predicting the spread across regions, due to the tendency of the neural network models to forecast overly-aggressive spread. As such, future work should address these data sparsity issues in order to increase the accuracy of CWD predictions across states. In conclusion, this study finds that while data sparsity is a pertinent issue, GNNs can still play a meaningful role in managing CWD and reducing its effects on both cervid and human populations.

Acknowledgments

Acknowledgments to Dr. B. Aditya Prakash and Shangqing Xu for their extensive support and guidance during the Data Science for Epidemiology course (CSE 8803 EPI) at Georgia Tech.

References

[1] Md Sohel Ahmed, Brenda J Hanley, Corey I Mitchell, Rachel C Abbott, Nicholas A Hollingshead, James G Booth, Joe Guinness, Christopher S Jennelle, Florian H Hodel, Carlos Gonzalez-Crespo, et al. 2024. Predicting chronic wasting disease in white-tailed deer at the county scale using machine learning. *Scientific reports* 14, 1 (2024), 14373.

[2] Qi Cao, Renhe Jiang, Chuang Yang, Zipei Fan, Xuan Song, and Ryosuke Shibasaki. 2023. Metapopulation graph neural networks: Deep metapopulation epidemic modeling with human mobility. *arXiv preprint arXiv:2306.14857* (2023).

[3] Centers for Disease Control and Prevention. 2024. Chronic Wasting Disease in Animals. https://www.cdc.gov/chronic-wasting/animals/index.html Accessed: 2025-10-07.

[4] Centers for Disease Control and Prevention. 2025. Where CWD Occurs. https://www.cdc.gov/chronic-wasting/data-research/index.html Accessed: 2025-10-07.

[5] Scott J Chiavacci. 2022. The economic costs of chronic wasting disease in the United States. *PLoS One* 17, 12 (2022), e0278366.

[6] Dane Erickson, Carson Reeling, and John G Lee. 2019. The effect of chronic wasting disease on resident deer hunting permit demand in Wisconsin. *Animals* 9, 12 (2019), 1096.

[7] Junkai Mao, Yuexing Han, and Bing Wang. 2024. MPSTAN: Metapopulation-Based Spatio-Temporal Attention Network for Epidemic Forecasting. *Entropy* 26, 4 (2024), 278. doi:10.3390/e26040278

[8] National Park Service. 2016. Chronic Wasting Disease. https://www.nps.gov/romo/learn/management/chronic-wasting-disease.htm. Last updated: March 31, 2016; Accessed: October 7, 2025.

[9] U.S. Census Bureau. 2024. County Adjacency File, 2024. https://www.census.gov/geographies/reference-files/time-series/geo/county-adjacency.2024.html. Accessed: 2025-11-03.

[10] Yufan Zheng, Wei Jiang, Alexander Zhou, Nguyen Quoc Viet Hung, Choujun Zhan, and Tong Chen. 2024. Epidemiology-informed Graph Neural Network for Heterogeneity-aware Epidemic Forecasting. arXiv:2411.17372 [cs.LG] https://arxiv.org/abs/2411.17372