Unraveling the Nexus of Illnesses and Heatwaves: Predictive Modeling for Early Warning Systems

Sai Nikhil Vangala*
svangala3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Suchet Sapre*
ssapre31@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Rigved Goyal*
rigvedgoyal@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

The alarming increase in the frequency and intensity of heat waves in recent years, fueled by the ongoing climate change, presents a substantial and urgent threat to public health worldwide. As global temperatures continue to rise, the need to predict, understand, and mitigate heat-related illnesses (HRIs) has become a paramount concern for scientists, policymakers, and healthcare professionals. This research project embarks on a multifaceted exploration of the intricate correlations between temperature and HRIs, with a primary focus on the Pacific Northwest region, a geographical area emblematic of the climate challenges faced by diverse communities. This study aims to reveal the nuanced dimensions of heat vulnerability, which can vary significantly from place to place. The ultimate goal is to contribute to the development of early warning systems and strategies that safeguard vulnerable populations not only in the Pacific Northwest but also in other regions, addressing the adverse effects of rising temperatures and heat-related health problems on a global scale.

CCS CONCEPTS

• **Computing methodologies → Modeling methodologies**;

KEYWORDS

Heat-related illnesses, Heatwaves, Predictive Modeling

1 INTRODUCTION

In recent years, the world has borne witness to a disconcerting surge in the frequency and intensity of heat waves, casting a looming shadow on the global landscape. This worrisome trend is undeniably linked to the pervasive issue of climate change and poses a substantial threat to public health. As global temperatures continue their inexorable ascent, the urgency to predict and mitigate the dire consequences of heat-related illnesses (HRIs) has become a paramount concern for scientists, policymakers, and healthcare professionals alike.

In response to this burgeoning challenge, our research project embarks on a multifaceted exploration, delving deep into the intricate web of correlations between temperature variations and the incidence of heat-related illnesses. Our primary focus centers on the Pacific Northwest region of the United States, which, although often recognized for its mild climate, is not immune to the mounting impacts of climate change. Here, we seek to unravel the intricate relationships between rising temperatures and the health of the populace, while simultaneously endeavoring to measure and compare these correlations across diverse geographic areas.

The aim of our study extends beyond mere observation; it aspires to shed light on the pressing inquiries that surround this intricate issue, offering valuable insights into the development of early warning systems and strategies to safeguard vulnerable populations. This research is not only timely but also essential. As climate change continues to shape our world, understanding the complex dynamics of temperature and health is a crucial step towards preserving the well-being of our communities. By focusing our attention on the Pacific Northwest and beyond, we aim to provide a foundation for evidence-based decision-making, the implementation of adaptive strategies, and the protection of those most at risk from the adverse effects of rising temperatures and heat-related health problems.

2 RESPONSE TO MILESTONE COMMENTS

In response to the provided feedback from the Milestone, we have addressed all of the suggestions in this final report. To begin with, we avoided direct screenshots for tables and also used unified caption format of figures as we provided all of the captions below each figure. In terms of the references, we now have 16 to have a more comprehensive literature. In addition, we have discussion the reference for XGBoost in the literature review. We have decided to go with LSTM instead of Transformer for final evaluation as we have found more reference material for it. Finally, we did not add analysis on regional differences in prediction results and performance because it was not in our problem definition and would not be able to present it in our paper due to the 8 page hard limit.

3 RELATED WORK

First, the paper titled "Machine and deep learning for modeling heat-health relationships" presented a robust approach by comparing six machine and deep learning models with three traditional statistical models to model heat-related mortality in Montreal, Canada. One of its key strengths is its comprehensive modeling approach, offering insights into the relative performance of various modeling techniques. However, its primary weakness is its limited geographic scope, which hinders the generalizability of its findings to regions with different climatic and demographic characteristics. To address this limitation, future research could expand the study to multiple locations. The short-term analysis focusing on daily mortality during summer months is another weakness, as it might not capture the long-term health impacts and cumulative effects of heatwaves. To address this, longer-term data can be considered and the health impacts beyond daily mortality patterns can be explored [1].

Second, the paper "Heatwave Damage Prediction Using Random Forest Model in Korea" is notable for utilizing big data to predict heat-related damages in South Korea. It effectively employs a random forest model, which can handle complex relationships and interactions between variables. However, its primary weakness lies in its limited geographical focus, which limits the generalizability of its findings. To address this, researchers could investigate the transferability of the model to other regions with different climate patterns and demographics. The use of data from 2015-2018 may not fully capture recent climate change trends, which is another weakness. To enhance the paper's relevance and robustness, researchers could consider incorporating more recent data. Additionally, discussing the practical implementation of the model in disaster response efforts would make the research more actionable and impactful [2].

The third paper, "Projection of heat wave mortality related to climate change in Korea" addressed the critical issue of climate change's impact on heatwave-related deaths using regression analysis. Its strength lies in providing a statistical basis for understanding the causal factors influencing heat-related deaths. Moreover, it considers multiple climate change scenarios and aging population scenarios, offering a comprehensive assessment of future risks. However, it has weaknesses in terms of generalizability, as it does not discuss whether similar patterns could be observed in other countries. To enhance its relevance, future research could explore the applicability of its findings to different regions. Additionally, the paper could benefit from discussing potential mitigation strategies or policies to reduce heat-related mortality, thereby offering more practical recommendations [3].

The fourth paper, "Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models" exceled in using machine learning techniques to forecast annual heatwave days in Iran. Its consideration of regional variations in performance adds valuable insights. However, it identifies ABR-DT as the best-performing model without an extensive comparison with other machine learning or statistical models, which is a notable weakness. To strengthen the paper, researchers could conduct a more comprehensive comparative analysis to validate the superiority of the chosen model. Additionally, addressing the interpretability challenges associated with decision trees and random forests would enhance the understanding of how predictors contribute to heatwave forecasting. The paper's Iran-centric focus may limit its applicability to other regions; hence, discussing the potential transferability of the approach to different geographical locations would broaden its relevance [4].

Next, the paper titled "A random forest model to predict heatstroke occurrence for heatwave in China " focused on predicting heatstroke occurrences in Chinese cities, considering both meteorological and socioeconomic parameters. Its strengths lie in addressing a pressing real-world issue and recognizing the multidimensional nature of heatstroke prediction. The paper's model outperforms traditional linear regression models, indicating the effectiveness of the proposed methodology. To improve the robustness of predictions, researchers could consider incorporating a more extended dataset. Additionally, the paper's regional focus on hot-temperature cities in China limits its generalizability. To address this limitation, future research could explore the model's applicability to other regions with varying climates and demographics. [5].

The paper "Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models" pioneered detailed spatial predictions of heat-related deaths within urban areas. Its strengths lie in its spatial resolution and novelty. However, the paper should address data generalizability issues by discussing the potential limitations of applying city-specific models to different regions. Validation methods for the model should also be explained.

The paper "Weekly heat wave death prediction model using zero-inflated regression approach" addresses weekly heat-related deaths in South Korea using a zero-inflated Poisson regression model. Its strength is its relevance to climate change and vulnerable populations. However, the paper should ensure data quality and provide clear explanations for the chosen statistical approach. Generalizability to other regions should be discussed [6].

The paper on "Weekly Heat Wave Death Prediction" addressed the urgent concern of heat-related fatalities in South Korea due to climate change and vulnerable populations. It employs a zero-inflated Poisson regression model to predict weekly heat-related deaths, considering factors like temperature, heatwave duration, and demographics. Its strengths include its relevance to public health and robust statistical approach. However, it depends on data quality, may be complex for non-experts, and lacks generalizability to other regions. Future research should explore data quality, enhance clarity, and discuss applicability to different locations [7].

The paper titled "The Effects of Summer Temperature and Heat Waves on Heat-Related Illness in Ningbo, China (2011-2013)" investigated the impact of extreme heat and heatwaves on heat-related illnesses. It used a distributed lag non-linear model (DLNM) to study the relationship between extreme temperatures and daily heat-related illnesses, finding that maximum temperature, rather than the heat index, was a better predictor of such illnesses, especially in the short term. The research identified six heatwaves during the study period, all associated with increased heat-related illnesses. Key findings include the substantial and delayed effects of recent heatwaves on health, a significant impact on severe heat diseases. While valuable, the research also highlights the need for broader geographic studies, longer-term perspectives, and discussions of strategies for addressing health risks associated with heatwaves to enhance its impact in public health initiatives [8].

The paper titled "Association between high temperature and heatwaves with heat-related illnesses: A systematic review and meta-analysis" significantly enriches our understanding of the impact of increased temperatures on human health. This comprehensive study systematically reviews and conducts a meta-analysis of a wide range of research findings, demonstrating a substantial increase in direct heat illness morbidity and mortality with each 1°C temperature rise, while also highlighting the vulnerability of specific populations and regions, such as individuals aged over 65 and those residing in subtropical climates. As part of the growing body of knowledge in the field, this paper underscores the critical need for preventative measures to address heat-related illnesses, particularly in the context of climate change, thus contributing to advancing our ability to predict and mitigate the health risks associated with extreme heat [9].

In the study "Machine Learning-Based Mortality Prediction for Hospitalized Heat-Related Illness Patients in Japan" is a significant addition. Leveraging data from a Japanese heat-related illness registry, the research developed and validated a mortality prediction model using logistic regression, support vector machine, random forest, and XGBoost. Notably, the models exhibited superior performance, with area under the precision-recall curve (AUPR) values ranging from 0.415 to 0.528, outperforming the conventional APACHE-II score. The consistently high area under the receiver operating characteristic curve (AUROC) values, surpassing 0.92 for all models, highlight the efficacy of machine learning, especially XGBoost, in enhancing mortality prediction for heat-related illnesses. This pioneering study not only improves prognostic accuracy but also holds promise for broader applications in healthcare contexts globally [10].

The paper "Social implementation and intervention with estimated morbidity of heat-related illnesses from weather data: A case study from Nagoya City, Japan" introduces novel frameworks to estimate heat-related illness cases in Nagoya City's administrative wards using 2014–2019 data. Employing derivation of estimation formulae and machine learning, both frameworks demonstrated impressive accuracy, with a daily residual estimation error below one person across 16 wards. Noteworthy is the finding that the daily working time average ambient temperature correlates better with ambulance-transported patients from outdoor sites than daily average or highest temperature. These frameworks not only enhance prediction precision but also provide insights for efficient ambulance allocation and public awareness strategies on hot days, aiming to curb heat-related morbidity effectively [11].

The paper "Approaches for estimating effects of climate change on heat-related deaths: Challenges and opportunities" addresses the complexities of forecasting health impacts in the face of shifting temperatures and increased heatwaves due to climate change and reviews three key approaches. Firstly, it explores historical weather–mortality relationships in the same region or a climatically similar location. Secondly, it evaluates adaptation using the minimum mortality threshold temperature, indicating the temperature with the lowest mortality rate. Lastly, it considers the impact of modifiers on temperature-mortality relationships, projecting effects based on plausible future parameter values in a specific city. While each approach sheds light on potential impacts, uncertainties persist. Nevertheless, projecting the future public health burden related to temperature effects remains crucial for informing public health and environmental planning in the face of climate change risks [12].

The study "Estimation of heat-related morbidity from weather data: A computational study in three prefectures of Japan over 2013–2018" addresses the increasing heat-related morbidity and mortality linked to global warming, emphasizing the need for accurate estimations to guide intervention and ambulance planning. Analyzing 95,137 ambulance transport cases in Japan in 2018, the research employs a computational technique to estimate daily peak core temperature elevation and water loss from 2013 to 2018 data in Tokyo, Osaka, and Aichi. The weighted sum of water loss and daily average ambient temperature over consecutive days proves more effective than conventional weather data, providing valuable insights for targeted intervention and ambulance planning in the context of heat-related morbidity [13].

The study "Mean radiant temperature – A predictor of heat related mortality" underscores the significance of mean radiant temperature (T_{mrt}) in health research, challenging the conventional use of air temperature (T_a) in analyzing weather's impact on mortality. T_{mrt}, a crucial meteorological parameter influencing human thermal comfort and heat load, is particularly valuable for assessing weather-related health impacts, especially in heat-related scenarios. Its direct correlation with urban geometry and surface material makes it a robust measure for identifying urban hotspots. Comparing the performance of models using T_a and T_{mrt} for daily mortality in Stockholm County, Sweden, the study reveals that T_{mrt} models better fit heat-related mortality, emphasizing the importance of considering T_{mrt} over T_a in health studies. Utilizing T_{mrt} allows for more accurate threshold determination for increased risks of heat-related mortality, enabling the identification of adverse weather conditions and heat-prone urban geometries. This information is crucial for implementing effective heat-warning systems and mitigating the harmful effects of heat stress [14].

The paper "Projecting Future Heat-Related Mortality under Climate Change Scenarios: A Systematic Review" explores heat-related mortality emerges as a critical public health issue. This systematic review, conducted in August 2010, scrutinizes existing methodologies for projecting future heat-related mortality under diverse climate change scenarios, drawing from peer-reviewed articles published in English from January 1980 to July 2010. The analysis of fourteen studies reveals a consensus that climate change is expected to significantly increase heat-related mortality. Emphasizing the pivotal role of scenario-based projection research, this review underscores its significance in assessing and managing potential climate change impacts on heat-related mortality for informed public health planning and effective intervention strategies [15].

Finally, The study "Projecting heat-related excess mortality under climate change scenarios in China" assesses the vulnerability of individuals and cities to climate change, projecting excess cause-, age-, region-, and education-specific mortality due to future high temperatures in 161 Chinese districts/counties. Utilizing 28 global climate models under two representative concentration pathways (RCPs), heat-related excess mortality is estimated to increase from 1.9% in the 2010s to 2.4% in the 2030s and 5.5% in the 2090s under RCP8.5. Additionally, the study examines the influence of population ageing on future heat-related mortality under five shared socioeconomic pathways (SSPs), revealing a 2.3- to 5.8-fold amplification in heat-related excess deaths. These findings offer crucial insights for shaping public health responses to mitigate climate change risks.

In summary, these papers contribute to our understanding of heat-related mortality modeling, but their strengths and weaknesses highlight the need for comprehensive, geographically diverse, and long-term studies in this crucial field. Addressing these limitations and building upon their strengths can further advance our ability to predict and mitigate heat-related health risks. Furthermore, the papers helped us formulate our methodology with the models we chose and define the gaps in research.

4 RESEARCH DEFINITION & INITIAL RESULTS

4.1 Problem Definition

This project is centered around heatwave-related excess illnesses and death prediction. The inherent predictive nature of this project implies that we will be taking data from the past and the present and using it to forecast into the future. However, to get to that stage, we are breaking our project down into two main tasks:

1) Comparing Correlations Between Temperature and Heat-Related Illnesses in Different Regions;
2) Using temperature data to predict Heat-Related Illnesses (HRIs) in the Pacific Northwest region, also known as region 10.

4.1.1 Comparing Correlations Between Temperature and Heat-Related Illnesses in Different Regions. This research task focuses on comparing the correlations between temperature data (independent variable X) and Heat-Related Illnesses (HRIs) in various regions (dependent variable Y). By analyzing temperature data and corresponding HRIs data in multiple regions, the study aims to identify regional variations in the relationship between temperature and HRIs. This comparative analysis will provide insights into how environmental factors interact with human health in different locations. The ultimate aim of this research task is to shed light on the regional differences in how temperature influences public health, which can inform region-specific public health policies and interventions.

4.1.2 Using temperature data to predict Heat-Related Illnesses (HRIs) in the Pacific Northwest region. This research aims to utilize temperature data as the primary variable (independent variable) to predict occurrences of Heat-Related Illnesses (HRIs) in the specific region of the Pacific Northwest (PNW). The objective is to analyze historical temperature data and correlate it with instances of HRIs, seeking to establish a predictive model. The research involves understanding the statistical relationships between temperature variations (a time-series, represented as $X_1, X_2, \ldots, X_n$) and the occurrences of HRIs (a time-series, represented as $Y_1, Y_2, \ldots, Y_n$). By exploring these correlations, the study aims to predict the likelihood of HRIs based on temperature fluctuations in the PNW region in the future. The ultimate goal of this research task is to enhance our ability to predict HRIs in the Pacific Northwest region by harnessing temperature data as a valuable tool for early warning systems and public health preparedness. To formalize this predictive task, given a particular time t, we will attempt to predict Y_{t+1} (the illness count one time-step into the future) based on past data, namely, $(X_1, ..., X_t)$ and $(Y_1, ..., Y_t)$. Note, however, we will only use $(Y_1, ..., Y_t)$ as our past data for the ARIMA model which we will describe later (due to its auto-regressive nature).

4.2 Data Collection

In order to create the correlations needed for this project, heat-related illness (HRI) data, as well as temperature data from different regions of the US is needed. This data is retrieved from the Center of Disease Control (CDC), specifically the Track Network Data API provided by the CDC. This REST API provides maximum temperature data at a daily granularity going back more than 10 years for every county in the continental US, as well as average temperature data over the course of a week for every week and county in the last 10 years. In addition, this API also groups states into 10 geographic regions and provides weekly and daily HRI data for each region. This regional/county level temperature and HRI data will make the model's predictions for HRI based on temperature robust and precise due to the small granularity of the data, which allows the model to train on small fluctuations from week to week in temperature and HRIs. The regions are composed of states within the continential United States in the follow grouping:

Region 1: ME, NH, VT, CT, MA, RI
Region 2: NY, NJ
Region 3: PA, MD, DE, VA, WV, DC
Region 4: $KY, NC, SC, TN, GA, FL, AL, MS$
Region 5: MN, WI, MI, IL, IN, OH
Region 6: AR, LA, OK, TX, NM
Region 7: IA, MO, KS, NE
Region 8: ND, SD, MT, WY, CO, UT
Region 9: AZ, CA, NV
Region 10: WA, OR, ID

This data was fetched using a REST API call to https://ephtracking.cdc.gov/DataExplorer/getCoreHolder/1237 for regional HRI data and https://ephtracking.cdc.gov/DataExplorer/getCoreHolder/1025 for county-level temperature data. These APIs return large JSON files with various metadata, and this metadata can be parsed and filtered into a CSV for training and testing. The CSV derived from this JSON contains the following data:

Regional Temperature/HRI Data				
County	**Avg. Temp.**	**Region**	**Illnesses**	**Week Of**
Autauga, AL	82.3	4	30	10/28/2023
Baldwin, AL	82.6	4	30	10/28/2023
Barbour, AL	81.2	4	30	10/28/2023
Bibb, AL	82.1	4	30	10/28/2023
Blount, AL	79.3	4	30	10/28/2023
Bullock, AL	81.1	4	30	10/28/2023
Butler, AL	82.7	4	30	10/28/2023
Calhoun, AL	78.7	4	30	10/28/2023
Chambers, AL	77.8	4	30	10/28/2023
Cherokee, AL	88.2	4	30	10/28/2023
...	...	...	...	...

Figure 1: Temperature/HRI data for regions and counties

4.3 Dataset Overview

The CDC provided us with a rich initial set of data to work with, however, prior to using it for any modeling or algorithmic purposes, we performed a few preprocessing steps. Firstly, we only obtained data from the CDC for 260 weeks prior to 10/28/2023. In other words, we collected data from the weeks of 11/10/2018 to 10/28/2023. Thereafter, we created a new column in our dataset to represent the week number. The week number for a particular week is defined as the number of weeks after 11/10/2018 it is. To clarify, 11/10/2018 is week 0, 11/17/2018 is week 1, ..., 10/21/2023 is week 258, and 10/28/2023 is week 259. This new column will make it easier to discuss the temporal nature of the data.

In addition, the COVID-19 pandemic occurred during the window over which we collected our data. To avoid any confounding affects caused by the pandemic, we decided to remove all data from our dataset between the weeks of 03/07/2020 and 03/06/2021 (inclusive).

Another important decision we made was regarding how to deal with the mismatch in granularity between the temperature and illness data. Specifically, we were able to obtain temperature data at the county level while we were only able to obtain illness data at the regional level. To resolve this discrepancy, for every week, we took the average of the temperatures of all the counties in each region. Thus, we are storing the average regional temperature along with illness count for every region for every week in our dataset. Figure 2 demonstrates how the first 20 rows of our dataset appear after these preprocessing steps.

Temperature/HRI data after preprocessing				
Region	**Avg. Temp.**	**Illnesses**	**Week Of**	**Week #**
1	61.855	16	10/28/2023	259
2	64.652	9	10/28/2023	259
3	69.195	13	10/28/2023	259
4	77.844	30	10/28/2023	259
5	64.359	6	10/28/2023	259
6	81.997	16	10/28/2023	259
7	69.820	0	10/28/2023	259
8	51.643	7	10/28/2023	259
9	68.514	18	10/28/2023	259
10	51.643	8	10/28/2023	259
1	58.146	4	10/21/2023	258
2	58.431	5	10/21/2023	258
...	...	...	...	...

Figure 2: Temperature/HRI data after preprocessing

After tabulating our data, we conducted some initial analysis to see how the trends in the average temperature are related to the regional illness count. The graph below shows this analysis for Region 10. We observed that increases in regional average temperature were accompanied by increases in regional illness count. In the next section of this paper, we will more formally outline this relationship.

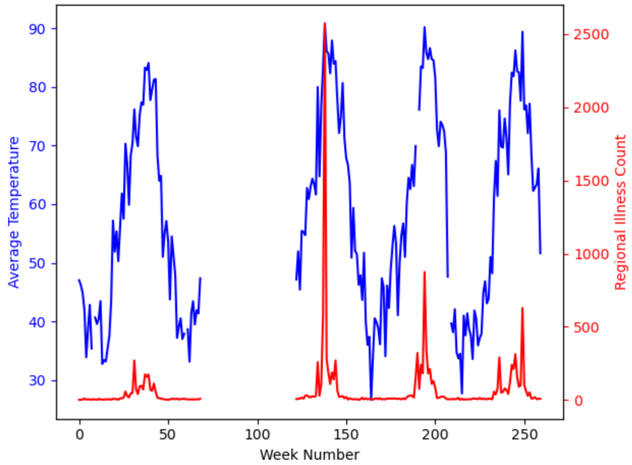

Figure 3: Average Temperature and Regional Illness Count vs. Week Number for Region 10

4.4 Initial Findings

In order to address task 1 in section 4.1, we decided to use the Pearson Correlation Coefficient (PCC) to measure how correlated each region's historical average temperature trend is to its illness trend. To perform this analysis, we first separated the dataset by region and then tabulated the appropriate average temperature and illness columns. PCC is a standard means of measuring linear correlation. We obtained PCC values for each of the 10 regions and our results are displayed in the table below.

PCC and temperature values for each region			
Region	**Avg. Temperature**	**PCC**	**Avg. HRIs**
1	57.281056	0.603907	39.14
2	59.710965	0.596408	28.80
3	65.553054	0.677653	61.67
4	73.732904	0.761961	93.74
5	59.148623	0.591785	44.79
6	76.285816	0.713821	135.23
7	64.171449	0.639717	116.47
8	56.544747	0.703789	57.55
9	70.703420	0.742930	122.34
10	58.070153	0.424426	58.41

Figure 4: PCC and temperature values for each region

From the above table, it is evident that Region 4 demonstrated the strongest correlation between average temperature and illness count while Region 10 demonstrated the weakest such correlation. It is also evident that regions that have higher average temperatures (which tend to be in the southern half of the US) and higher volume of HRIs also tend to have stronger temperature/HRI correlation, as indicated by PCC v.s average temperature/HRI values in the figure above. There could be several reasons for this, but one reason this could be the case is because in cooler regions, even as temperatures rise towards their peak during warmer weeks, temperatures don't reach high enough to consistently cause HRIs since humans have a certain resiliency to heat, and thus HRIs happen more sporadically and inconsistently. However, in warmer regions, as temperatures reach their peaks during warmer weeks, the heat is high enough to more consistently break this resiliency, and thus the number of HRIs tend to coincide with the rise in temperatures more strongly in these regions.

5 PROPOSED METHODOLOGY

5.1 Advancing Beyond the State-of-Art

The project's strategy for predicting heatwave-related excess illnesses and death combines classical time series models such as ARIMA with modern machine learning models like XGBoost and deep learning models like LSTM, forming a versatile and comprehensive approach. By using ARIMA as a baseline, the project acknowledges and incorporates temporal dynamics, capturing both linear and complex nonlinear patterns through the hybridization of models. The choice of XGBoost, known for its effectiveness in time-series classification, and the application of a grid search for hyperparameter tuning demonstrate a commitment to preventing overfitting and obtaining optimal results. Additionally, the incorporation of LSTM, specifically designed for time series data, involves a grid search for hyperparameters, indicating a dedication to fine-tuning the model for dataset-specific characteristics. The comparison and contrast of ARIMA, XGBoost, and LSTM provide nuanced insights into the suitability of different models for the prediction task, addressing whether a more complex model is necessary. The implementation with specialized libraries like xgboost and scalecast reflects an optimization-focused approach. Furthermore, there has not been effective methodological research studies done on the different regions in the United States as it has been done in China and Japan as stated by the literature review. The emphasis on the Pacific Northwest region enables for a much more impactful study. Overall, this methodology, with its careful model selection, hyperparameter tuning, and hybridized approach, is superior to state-of-the-art methods.

5.2 Formal Model Description

In order to better grasp the models and tools that we will outline in this section, we will provide a formal mathematical description along with our implementation plan. The models that we will use fall into three main classes: classical time series models, supervised machine learning (ML) models, and deep learning-based models.

It is important to note that we will be using the same formalism outlined in section 4.1. That is, for a particular region R, the average temperature data is a time series which is represented as $X_0, X_1, \ldots, X_{206}$ and the illness count is also a corresponding time series which is represented as $Y_0, Y_1, \ldots, Y_{206}$. Note that there are 207 values in these series instead of 260 values because we removed data between the weeks of 03/07/2020 and 03/06/2021 (inclusive) to account for the pandemic.

5.2.1 Classical Models: ARIMA. Based on our literature review, one of the most common groups of classical time series models is the ARIMA family. The ARIMA model is a combination of both the autoregressive and moving average approaches. The autoregressive component of the ARIMA model computes a linear combination of a few of the past values of the independent variable. Additionally, the moving average component of the ARIMA model computes a linear combination of the past forecasting errors. ARIMA models are parameterized by three variables: p, d, q. The variable p defines the number of past values that are accounted for in the autoregressive component. The variable d defines the number of differences of the time series data that are taken. These differences are taken to ensure that the time series data is stationary or in other words that there are no trends, seasonality, etc. Finally, the variable q defines the number of past forecasting errors that are accounted for in the moving average component. There are numerous methods that we can use to determine how to best set these variables in order to fit an ARIMA model. For instance, the Augmented Dickey-Fuller test can be used to calibrate the d value. These methods along with a overarching grid-search based approach will be used to determine the best p, d, q values with which to fit the ARIMA model. Recall that the ARIMA model is autoregressive in nature and as a result it is designed for single-view time-series data. As a result, we can't include historical average temperature data when using this model (we can only use historical illness count to forecast itself). This means that it will serve as a baseline for the supervised ML and deep learning-based models that we will also use.

5.2.2 Supervised ML Models: XGBoost. The XGBoost supervised ML model is similar to the Random Forest model in that it makes a final classification decision based on a collection of weak learners (trees). However, it leverages the gradient boosting framework to produce consistently better results than comparable Random Forest models. XGBoost is commonly used for time-series related classification tasks. We plan to use the `xgboost` Python library in order to implement our classifier. We will experiment with the number of estimators our XGBoost models will use such that we both prevent over-fitting and still obtain locally optimal results.

5.2.3 Deep Learning-Based Models: LSTM. We plan to also use a Long-Short Term Memory (LSTM) type of RNN as it is highly performant on time series data. We will run a grid search on its hyperparameters (such as the learning rate) in order to find the network with the best fit. Comparing and contrasting both the XGBoost and LSTM models on this dataset will provide insight into whether a "deeper" model is needed for this prediction task. With respect to implementation, there are numerous Python libraries that can be used to implement LSTMs. After an initial search, we plan to use TensorFlow to construct LSTMs designed specifically for time-series classification tasks.

5.3 Evaluation

We will use root mean squared error (RMSE) in order to evaluate the predictive capabilities of the ARIMA, XGBoost, and LSTM models. This error metric is typically used to evaluate time series prediction tasks. We will test our method by first dividing up our data into train, test, and validation groups. Thereafter, we will evaluate our models only on the validation set to ensure unbiased results. Success of our methodology will be measured by comparing our error values to those obtained by other researchers (after accounting for differences between the datasets and normalization practices).

6 EXPERIMENTS/RESULTS

6.1 Experimental Questions

The experiments aim to answer several questions related to the relationship between average temperature and regional illnesses as follows: What is the nature of the relationship between average temperature and the occurrence of regional illnesses? To what extent can average temperature be used as a reliable indicator for predicting the prevalence of regional illnesses? Does the model effectively capture seasonal variations in regional illnesses based on fluctuations in average temperature? Can the ARIMA model, which is inherently univariate, accurately model regional illness trends? How does the baseline ARIMA model compare to the more complex models that we are using such as XGBoost and LSTM? How accurately does the XGBoost model predict regional illness levels based on average temperature, as indicated by the achieved RMSE value? Which model has the best performance?

6.2 ARIMA Model

The ARIMA model will serve as our baseline model of comparison. We focused specifically on Region 10 (as is the case for both the XGBoost and LSTM models). The first 80% of the dataset was used for training and the remaining 20% of the dataset was used for testing. We used a forecast window of 1 ($W = 1$) which means that we used the ARIMA model to forecast regional illness counts one week into the future. A rolling forecast-based model was used to perform the fitting. That is, we sequentially retrain our ARIMA model on new data as it is made available. For every regional illness count that we forecast, we add the ground truth observation corresponding to that prediction to our training set.

As mentioned in our proposed methodology, ARIMA models are parameterized by three variables: p, d, q, We implemented a grid search to determine which set of these three parameters would yield the best ARIMA model for region 10. We evaluated all ARIMA models with parameters in the following search space: $(p, d, q) \in [0, 1, 2, 4, 6, 8, 10] \times [0, 1, 2] \times [0, 1, 2]$. The best model had the parameters $(p, d, q) = (10, 0, 1)$ and yielded an RMSE value of 102.0521. The graph below shows the actual and predicted regional illness count values for the ARIMA model with $(p, d, q) = (10, 0, 1)$ for Region 10. The ARIMA model is clearly able to capture the general (seasonal) trends in the regional illness count. However, we also observed that the model is unable to account for large deviations in the data (e.g. between August and September 2023). This is expected as the ARIMA model is a fairly simplistic time-series model.

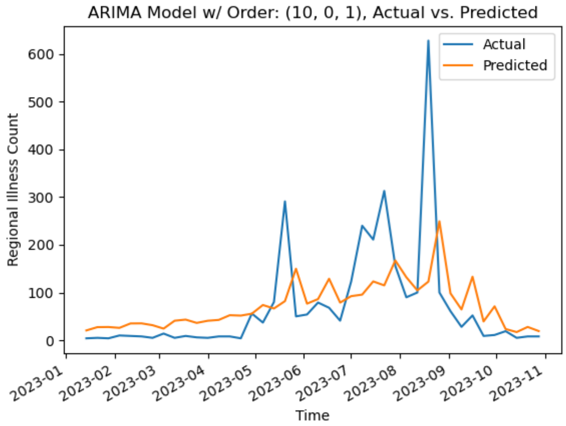

Figure 5: The visual representation of the model's predictions against the actual values

6.3 XGBoost Model

The XGBoost algorithm is used to develop a predictive model for regional illnesses, focusing specifically on Region 10. The dataset is split into training and testing sets, with 80% used for training and 20% for testing. We used a forecast window of 1 ($W = 1$) which means that we used the XGBoost model to forecast regional illness counts one week into the future. A rolling forecast-based model was used to perform the fitting. That is, we sequentially retrain our XGBoost model on new data as it is made available. For every regional illness count that we forecast, we add the ground truth observation corresponding to that prediction to our training set.

Evaluation of the model's performance is conducted on the test set, utilizing the Root Mean Squared Error (RMSE) as a metric. A visual representation of the model's predictions against actual values for the last weeks of the dataset is generated. his research contributes to the understanding of how machine learning techniques, specifically XGBoost, can effectively model the relationship between environmental factors, represented by average temperature, and the incidence of regional illnesses. The importance of hyperparameter tuning is highlighted through the grid search process, demonstrating improved predictive accuracy in the final model.

The Root Mean Squared Error (RMSE) from the XGBoost model for region 10 is 57.46. The value indicates that, on average, the model's predictions deviate from the actual values by approximately 57.46 units of the "Regional Illnesses" variable. The graphic below showcases the XGBoost Model with Grid Search of Actual values vs Predicted values. It is easy to notice that the plot lines align for the most part in terms of pattern. It is important to note that there a wide range of factors that contribute to heat related illnesses including patient's personal characteristics. The model's predictions are reasonably close to the actual values considering that temperature is the only factor that was considered.

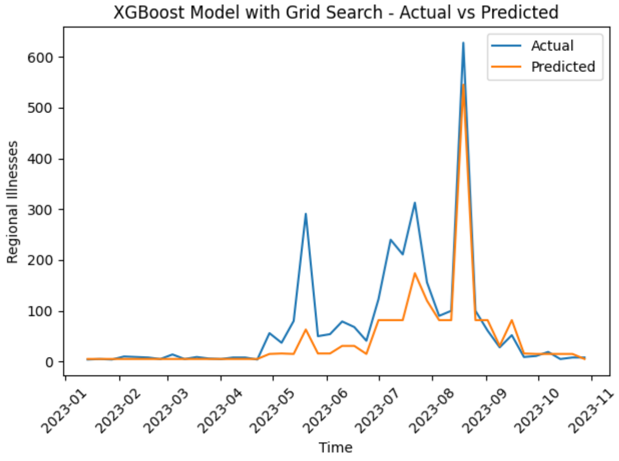

Figure 6: The visual representation of the model's predictions against the actual values

6.4 LSTM Model

The LSTM model, developed using Tensorflow, uses historic sequential temperature data to predict HRIs. For this experiment, this model was trained and tested on temperature/HRI data from Region 10 as tracked by the CDC. For each week, the LSTM model uses not just the temperature data for the current week, but from the last n weeks from the current week in sequence and trains with it against the number of HRIs from the current week. In order to train this model, preprocessing was carried out to split data into features (temperature) and a target (HRIs). Then, additional preprocessing turned the original 2D shaped feature data (260x1 weeks x column (temperature)) into 3-D shaped feature data (260x30x1 weeks x sequence length x column (temperature)). This means that for each week, temperature data from the last 30 weeks were considered in sequence to predict that week's number of HRIs. 30 weeks was considered for the sequence length as larger sequences tended to result in, albeit marginal, improvements in prediction performance. Finally, the preprocessing was finished by splitting the data into train and test data, with an 80%/20% split, respectively.

Once the data was preprocessed it was used to train the LSTM model. In order to calibrate the model, a grid search was conducted on various values for several hyperparameters such as units, dropout rate, learning rate, and activation method. The evaluation metric used to determine which combination of hyperparameters would yield the best prediction results was Root Mean Squared Error (RMSE). After calibrating hyperparameters, the best RMSE achieved by LSTM was 87.387, indicating that on average, the model's predicted number of HRIs differs from the actual number of HRIs by 87.387 over the 52 week period over which this model's predictions were made. As seen by the graph comparing predicted v.s actual HRI values, this is a reasonable margin of error, as predictions tend to generally track with spikes in HRIs (with the exception of an abnormal, 1 week spike around 2023-09).

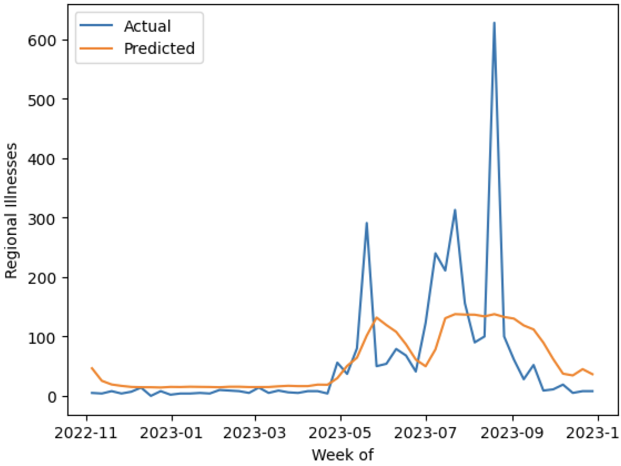

Figure 7: The visual representation of the LSTM model's predictions against the actual values

7 CONCLUSION & FUTURE WORK

As mentioned in the problem definition, there are two main tasks addressed in this paper:

1) Comparing Correlations Between Temperature and Heat-Related Illnesses (HRIs) in Different Regions;
2) Using temperature data to predict HRIs in the Pacific Northwest region.

With respect to the first task, we observed that Region 4 demonstrated the strongest correlation between average temperature and illness while Region 10, Pacific Northwest, demonstrated the weakest such correlation. In addition, Figure 4 indicates that regions with higher average temperatures tend to exhibit stronger correlations between average temperature and illness. Overall, completing this first task allowed us to understand that temperature does indeed have an affect on regional illness count.

We approached the second task by constructing three models: ARIMA, XGBoost, and LSTM. The main goal of each of these models was to accurately forecast the regional illness count one week into the future using both past temperature and illness data. Each of these models was evaluated based on their RMSE value on the test dataset (the latter 20% of the dataset). A gridsearch on the parameter-space was used for all three models to determine the optimal parameter choice. We obtained RMSE values of *102.0521, 57.46*, and *87.387* for the ARIMA, XGBoost, and LSTM models, respectively. These results indicate that the XGBoost model performed the best. This is could be because the XGBoost model was able to account for the larger deviations in HRIs for Region 10 better than the LSTM model. In addition, LSTM bases its predictions off of historic temperature data. This means that if in a given week, an abnormal temperature spike occurs, the historic temperature data will balance out the abnormality and thus, the model likely won't predict a spike in HRIs. In other words, LSTM isn't as great with outlier data as XGBoost, which can be seen in the predictive plots for both of these models. We observed that the ARIMA model performed the worst. This was expected as the ARIMA model served as our baseline and didn't utilize the Region 10 temperature data (because it is uni-variate in nature).

In terms of future work for this project, there are three main avenues which we believe could yield fruitful results. Firstly, to expand upon the LSTM, exploring transformer and ensemble-based models could help bolster accuracy further by incorporating attention and information from multiple classifiers, respectively. Secondly, throughout our experiments, we kept our forecasting window at 1. In the future, we plan to increase this window to determine whether our models can accurately predict the number of HRIs farther into the future. Finally, our paper focuses on using only temperature data for each region. In the future, we would look into using more weather-related data such as wind speeds and humidity to inform our HRI forecasts. Using other weather-related data could reduce some of the errors exhibited by our models.

ACKNOWLEDGMENTS

We would like to thank Leo Zhao along with Dr. Prakash for providing the idea for this project proposal.

REFERENCES

[1] Boudreault, J., Campagna, C. and Chebana, F. (2023) 'Machine and deep learning for modelling heat-health relationships', *Science of The Total Environment*, 892, p. 164660. doi:10.1016/j.scitotenv.2023.164660.

[2] Park, M. et al. (2020) 'Heatwave damage prediction using Random Forest model in Korea', *Applied Sciences*, 10(22), p. 8237. doi:10.3390/app10228237.

[3] Kim, D.-W. et al. (2015) 'Projection of heat wave mortality related to climate change in Korea', *Natural Hazards*, 80(1), pp. 623–637. doi:10.1007/s11069-015-1987-0.

[4] Asadollah, S.B. et al. (2021) 'Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models', *Stochastic Environmental Research and Risk Assessment*, 36(7), pp. 1959–1974. doi:10.1007/s00477-021-02103-z.

[5] Wang, Y. et al. (2019) 'A random forest model to predict heatstroke occurrence for heatwave in China', *Science of The Total Environment*, 650, pp. 3048–3053. doi:10.1016/j.scitotenv.2018.09.369.

[6] Kim, Yesuel and Kim, Youngchul (2022) 'Explainable heat-related mortality with random forest and Shapley additive explanations (SHAP) models', *Sustainable Cities and Society*, 79, p. 103677. doi:10.1016/j.scs.2022.103677.

[7] Kim, D.-W. et al. (2018) 'Weekly Heat Wave Death Prediction model using zero-inflated regression approach', *Theoretical and Applied Climatology*, 137(1–2), pp. 823–838. doi:10.1007/s00704-018-2636-9.

[8] Bai, L. et al. (2014) 'The effects of summer temperature and heat waves on heat-related illness in a coastal city of China, 2011–2013', *Environmental Research*, 132, pp. 212–219. doi:10.1016/j.envres.2014.04.002.

[9] Faurie, C. et al. (2022) 'Association between high temperature and heatwaves with heat-related illnesses: A systematic review and meta-analysis', *Science of The Total Environment*, 852, p. 158332. doi:10.1016/j.scitotenv.2022.158332.

[10] Hirano, Y. et al. (2021) 'Machine learning-based mortality prediction model for heat-related illness', *Scientific Reports*, 11(1). doi:10.1038/s41598-021-88581-1.

[11] Nishimura, T. et al. (2021) 'Social implementation and intervention with estimated morbidity of heat-related illnesses from weather data: A case study from Nagoya City, Japan', *Sustainable Cities and Society*, 74, p. 103203. doi:10.1016/j.scs.2021.103203.

[12] Kinney, P.L. et al. (2008) 'Approaches for estimating effects of climate change on heat-related deaths: Challenges and opportunities', *Environmental Science & Policy*, 11(1), pp. 87–96. doi:10.1016/j.envsci.2007.08.001.

[13] Kodera, S. et al. (2019) 'Estimation of heat-related morbidity from weather data: A computational study in three prefectures of Japan over 2013–2018', *Environment International*, 130, p. 104907. doi:10.1016/j.envint.2019.104907.

[14] Thorsson, S. et al. (2014) 'Mean radiant temperature – a predictor of heat related mortality', *Urban Climate*, 10, pp. 332–345. doi:10.1016/j.uclim.2014.01.004.

[15] Huang, C. et al. (2011) 'Projecting future heat-related mortality under climate change scenarios: A systematic review', *Environmental Health Perspectives*, 119(12), pp. 1681–1690. doi:10.1289/ehp.1103456.

[16] Jun Yang et al. 2021. Projecting heat-related excess mortality under climate change scenarios in China. *Nature Communications* 12, 1 (2021). DOI:http://dx.doi.org/10.1038/s41467-021-21305-1