New Jersey Institute of Technology

Digital Commons @ NJIT

Theses

Theses and Dissertations

12-31-2019

Predictive modeling of influenza in New England using a recurrent deep neural network

Alfred Amendolara
New Jersey Institute of Technology

ABSTRACT

PREDICTIVE MODELING OF INFLUENZA IN NEW ENGLAND USING A RECURRENT DEEP NEURAL NETWORK

by
Alfred Amendolara

Predicting seasonal variation in influenza epidemics is an ongoing challenge. To better predict seasonal influenza and provide early warning of pandemics, a novel approach to Influenza-Like-Illness (ILI) prediction was developed. This approach combined a deep neural network with ILI, climate, and population data. A predictive model was created using a deep neural network based on TensorFlow 2.0 Beta. The model used Long-Short Term Memory (LSTM) nodes. Data was collected from the Center for Disease Control, the National Center for Environmental Information (NCEI) and the United States Census Bureau. These parameters were temperature, precipitation, wind speed, population size, vaccination rate and vaccination efficacy. Temperature was confirmed as the greatest predictor for ILI rates, with precipitation providing a small increase in predictive power. After training, the model was able to predict ILI rates 10 weeks out. As a result of this thesis, a framework was developed that may be applied to weekly ILI tracking as well as early prediction of outlier pandemic years.

PREDICTIVE MODELING OF INFLUENZA IN NEW ENGLAND USING A RECURRENT DEEP NEURAL NETWORK

by
Alfred Amendolara

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Biology

Federated Biological Sciences Department

December 2019

APPROVAL PAGE

PREDICTIVE MODELING OF INFLUENZA IN NEW ENGLAND USING A RECURRENT DEEP NEURAL NETWORK

Alfred Amendolara

Dr. Eric Fortune, Thesis Advisor Date
Associate Professor of Biological Sciences, New Jersey Institute of Technology

---

Dr. Horacio Rotstein, Committee Member Date
Professor of Biological Sciences, New Jersey Institute of Technology

---

Dr. Kristen Severi, Committee Member Date
Assistant Professor of Biological Sciences, New Jersey Institute of Technology

BIOGRAPHICAL SKETCH

Author: Alfred Amendolara

Degree: Master of Science

Date: December 2019

Undergraduate and Graduate Education:

- Master of Science in Biology,
  New Jersey Institute of Technology, Newark, NJ, 2019
- Bachelor of Science in Biological Sciences,
  Fordham University, Bronx, NY, 2017

Major: Biology

Dad, I miss you. I hope you like it.

ACKNOWLEDGMENT

First and foremost, I would like to express my deepest gratitude to Dr. Fortune for his support and mentorship. The advice he has given me has been invaluable and the time spent in his lab has shaped me substantially. I very much appreciate the time that Dr. Rotstein and Dr. Severi have given me, listening and providing feedback as well as sitting on my committee. I am also grateful to Dr. Bucher for providing excellent guidance in this program. I would also like to recognize the assistance of Ms Roach, who made dealing with administrative issues so much easier.

Many thanks to Mohammad Farooq and Patrick Janeczko for reading and editing my thesis. Special thanks to Brielle Burns for putting up with my shenanigans and supporting me throughout the process. Also, thank you to all the friends at NJIT and Rutgers that made my time here memorable. Finally, I want to thank my Mom for her unwavering support and belief in me. I would not have been able to make it this far without her.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Influenza virus is responsible for a recurrent, yearly epidemic in most temperate regions of the world. According to the CDC, in the 2017-2018 season alone, influenza virus was responsible for 79,000 deaths and nearly 1 million hospitalizations [4]. The disease burden of influenza is substantial. As a result of viral mutation, as well as a variety of climate factors, seasonal trends can shift radically year to year. This necessitates the development of a new vaccine for each season. Despite the combined efforts of scientists around the world, vaccine efficacy is variable. In some years effectiveness has been estimated as low as 20% [106, 10]. Additionally, severe pandemics can occur with little warning. Most recently, the 2009 Swine Flu caused an unusually long and deadly flu season [73]. Modeling and forecasting of influenza is critical for predicting pandemics such as this.

This thesis will examine the structure, pathology, epidemiology, and evolution of the influenza virus in order to apply machine learning techniques to produce a forecasting model. First the structure of an individual influenza virus will be discussed, followed by the genome, important protein components and viral replication. Then the evolution of the virus. Next, the mechanisms of transmission and infection, the clinical signs and the immune response will be examined. Finally, this introduction will end with an overview of influenza epidemiology and applied modeling.

The goal of this paper is to produce an effective predictive model that will shed light on the factors that impact influenza seasonality as well as provide a functional predictive model for real-time flu forecasting. Variables including temperature, precipitation, wind speed, and vaccination rates were added to a deep recurrent neural network. The network was built on TensorFlow 2.0 Beta using the Keras API. An effective architecture was developed that provided robust predictions 1 week, 2 weeks, and 3 weeks out, as well as a framework for recurrent predictions that could continue on past several months.

CHAPTER 2

BACKGROUND

2.1 Structure and Replication

Since the discovery of the influenza virus, all the major components of the virus have been examined. Given the seasonal nature of influenza and the potential for major pandemics, there has been extensive research to elucidate viral structure and proteins. This has allowed scientists to choose targets for antivirals. Additionally, it has allowed epidemiologists to identify many viral subtypes. One example is the H1N1 strain that was responsible for the 2009 Swine-Flu Pandemic.

This section will summarize the basic structure of an influenza virus, the structure of the viral genome, the function of important proteins, and the replication of the virus.

2.1.1 Morphology

Influenza belongs to the viral family Orthomyxoviridae. There are four subtypes of influenza; A, B, C, and D. Influenza A is the most common type seen in humans and will be the primary focus of this section as it accounts for the majority of seasonal outbreaks and is more heavily studied. Influenza B is also regularly seen in humans, but to a lesser extent than influenza A. Influenza C is rarely seen in humans and does not generally contribute to seasonal influenza outbreaks. Influenza D does not occur in humans and will be ignored completely [3, 63]. Unless otherwise noted, we will be discussing influenza A.

Influenza is an enveloped, spherical negative-sense RNA virus approximately 100nm in diameter. It can also occasionally be filamentous. The influenza virus has three major structural components: an envelope, a matrix directly beneath the lipid bilayer, and ribonucleoproteins (RNP) in the center of the virus [63].

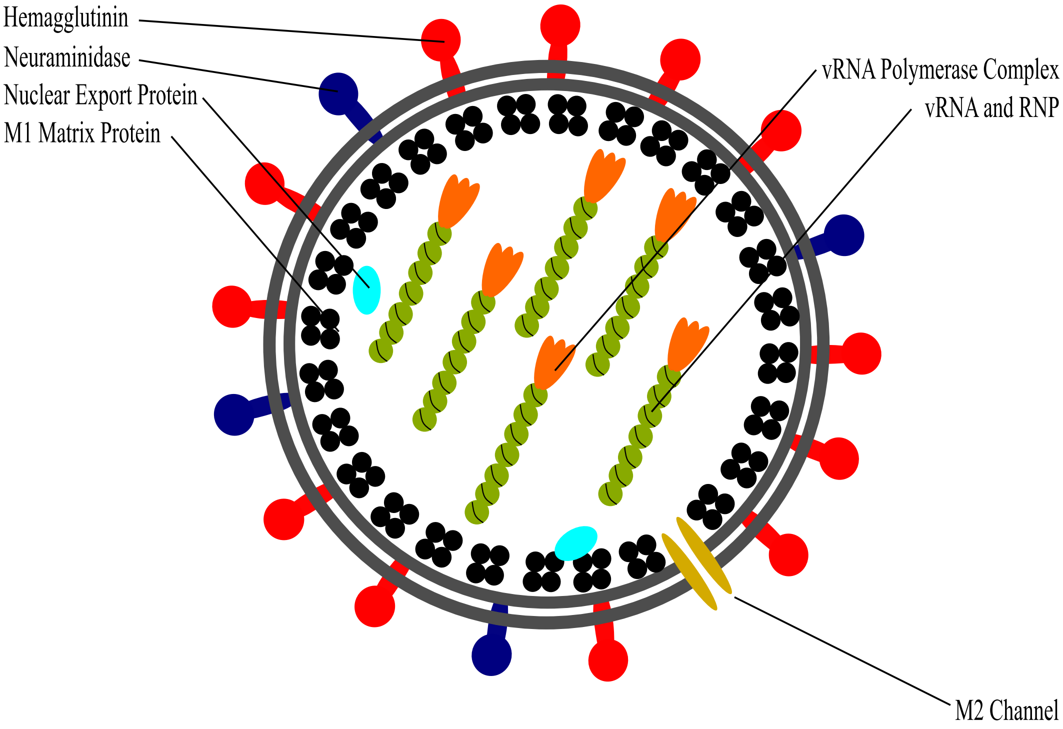

Figure 2.1 Diagram of influenza virus.

The outer envelope is composed of a lipid bi-layer derived from the host cell membrane. Lipid rafts containing neuraminidase (NA) float in a sea of hemagglutinin (HA). Hemagglutinin, which makes up about 80% of membrane bound protein, and neuraminidase, which makes up about 20%, are both important viral proteins necessary for infection and replication [65, 59, 77]. M2, a viral proton pump, is also embedded in small numbers in the envelope. The specifics of each of these proteins will be discussed in more detail later. The next major structural component is the matrix. The matrix exists directly beneath the envelope and serves to anchor NA, HA, and RNP inside the virus. The matrix is also involved in viral budding. It is made up of the matrix protein M1.

The inner most component of an influenza virus are the RNP. These are protein-RNA complexes that contain the genetic material necessary for viral propagation as well as various proteins. RNP are helical structures. In addition to the viral RNA, some nuclear export proteins (NEP) are present, as well as basic polymerase 1 (PB1), basic polymerase 2(PB2), and acidic polymerase (PA) [65, 64, 77].

2.1.2 Genome

The genome of the influenza virus contains 8 segments of single stranded RNA. Segment 1 encodes the various polymerase proteins and segments 2 and 3 encode proteins involved in virulence modulation [101]. In influenza A viruses, the focus of this section, segments 4, 5 and 6 code for hemagglutinin, nucleocapsid protein and neuraminidase [101, 46]. Nucleocapsid proteins are bound to viral RNA and play a role in viral genome replication [46]. Segments 7 and 8 code M2 along with nuclear export protein and structural proteins [49]. These coding regions are flanked on both 5' and 3' ends by non-coding regions [46]

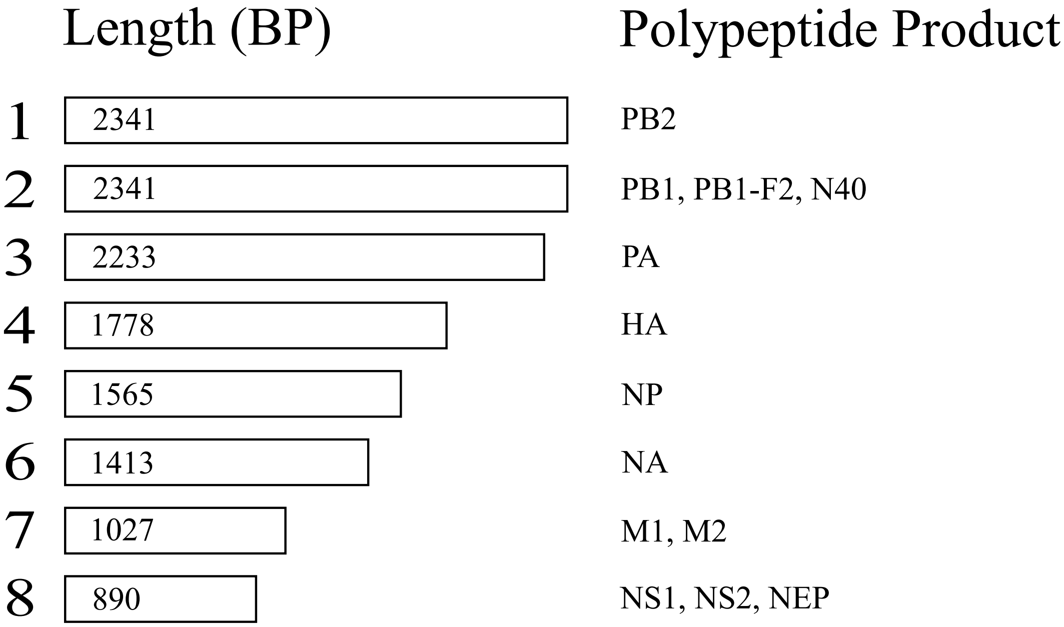

Figure 2.2 Genome of influenza A virus. This illustration shows the length of each viral RNA segment in base pairs (BP) as well as the polypeptide product. The genome of influenza B is similar in layout, but with different segment lengths [46].

The genome resides in the center of the virus particle. It is found in RNP complexes with RNA polymerases and NP. Viral RNA (vRNA) is associated with NP in strands. The genome is replicated by polymerases PB1, PB2, and PA [46, 102]. First, the RNP complexes are into the cytoplasm of an infected cell. Then, within the nucleus of the cell, viral polymerases replicate the vRNA [102]. This replication peaks approximately two hours after infection and begins to decrease at approximately 3 hours after infection [75, 83]. After replication, vRNA joins with NP and forms RNP complexes. These are transported out of the nucleus by viral NEP along with host factors. They then move toward sites of budding where they will be incorporated into new virus particles [69].

2.1.3 Hemagglutinin, Neuraminidase and M2

Hemagglutinin, Neuraminidase and M2 the the three surface proteins of the infleunza virus. Hemagglutinin is the major surface protein of influenza viruses. It is primarily responsible for binding sialic acid to allow entry into host cells [86]. It is a trimer composed of triple stranded α-helices and anti-parallel β-sheets [100]. A highly conserved region on the top of the HA molecule forms a depression where sialic acid bonds to. Specifically, there are two major linkages to sialic acid that HA forms. These are the α(2,3)-Gal and the α(2,6)-Gal linkages. Human viruses bind to α(2,6) while avian viruses bind to α(2,3) [86]. Because of these differences in binding preference, mutations must occur to allow inter-species transmission. Swine have both kinds of receptors and provide a reservoir for these kinds of mutations. Generally antigenic shift is prominent in the exterior, binding sections of HA, while distal sections tend to be retained [99, 98]. In general, blocking HA using anti-HA antibodies effectively neutralizes viral infectivity [99].

The other major surface protein found on influenza viruses is neuraminidase. It is a tetramer composed of four identical sub-units [59, 94]. The main role of neuraminidase is believed to cleavage sialic acid to release newly formed virus particles from the cell surface after budding[58]. It also may allow the virus to escape capture in mucus through the same mechanism. Neuraminidase stalk length may play a role in virulence as well, differing lengths have been shown to have differing levels of infectious activity. This may be a result of occlusion of the active site if the stalk is too short [66]. Additionally, neuraminidase ay be involved in infection of host cells. Its active site is also highly conserved [7].

The last of the surface proteins is M2, an ion channel. It is an integral membrane protein formed by a di-sulfide linked homotetramer [51, 88]. It tends to be present in very low quantities, only 15-20 molecules may be on the surface of any given virus particle [104]. Despite its low concentration compared to neuraminidase and hemagglutinin, it is essential for replication as it facilitates uncoating. It is also responsible for adjusting pH in the Golgi during viral replication [50, 79]. M2 may also inhibit P58, aiding in immune response evasion [32].

2.1.4 Viral Replication

Viral replication begins after a viron binds to the sialic acid receptors of a target cell. HA in particular is responsible for the binding of sialic acid (2.3.1). This triggers receptor mediated endocytosis and the viral particle enters the host cell. In this stage, M2 proton channels are responsible for lowing the pH of the endosome to induce conformational changes in HA. Lowering the pH also prompts the release of vRNPs into the cell [78]. These vRNPs can then enter the nucleus and begin replication (2.3.3).

The various proteins that make up a RNP each have nuclear localization signals that allow them to make use of cellular mechanisms for entry into the nucleus. Once the a RNP is transported into the nucleus, viral RNA makes use of many existing cellular mechanisms for replication. The negative sense vRNA is replicated by viral RNA dependant RNA polymerases [78].

Once replicated, negative sense vRNA is exported through nuclear pores to form completed RNPs. Proteins necessary to form the RNPs are produced in the cytoplasm. Surface proteins are produced in the ER and Golgi (2.3.4). Influenza makes use of the host cell membrane to form its envelope. Once completed surface proteins have been transported to the host membrane, RNPs are localized and bud off in newly formed viral particles (2.3.5) [63, 78]. From here, viral particles can continue the cycle of infection.

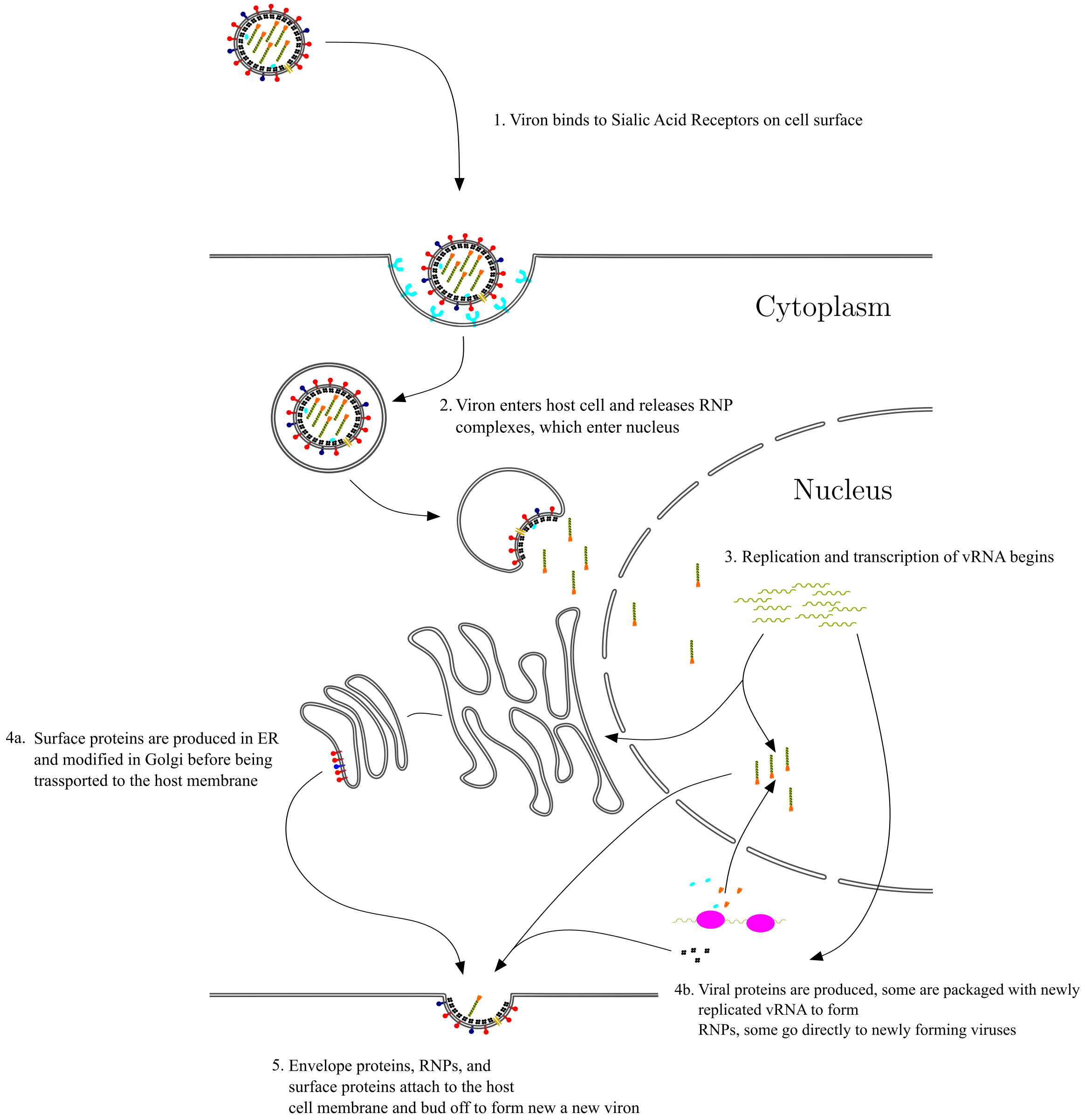
1. Viron binds to Sialic Acid Receptors on cell surface
Cytoplasm
2. Viron enters host cell and releases RNP complexes, which enter nucleus
Nucleus
3. Replication and transcription of vRNA begins
4a. Surface proteins are produced in ER and modified in Golgi before being trassported to the host membrane
4b. Viral proteins are produced, some are packaged with newly replicated vRNA to form RNPs, some go directly to newly forming viruses
5. Envelope proteins, RNPs, and surface proteins attach to the host cell membrane and bud off to form new a new viron

Figure 2.3 Influenza virus replication.

2.2 Evolution

Influenza has almost certainly existed through human history and the earliest cases may have been recorded by Hippocrates [57]. The disease was certainly known. However, it was not until the 1930's that the virus was isolated in pigs [84]. Even with the advent of modern medicine, we are subject to regular, seasonal epidemics and even occasional worldwide pandemics. Despite the creation of vaccines that wiped out diseases like small-pox and polio and the advent of antiviral medication, influenza has remained a constant burden. Influenza's rapid mutation is largely responsible for its ability to elude eradication [17].

2.2.1 Mutation and Antigenic Drift

Influenza virus has a high mutation rate, approximately $1.5x10^-5$ mutations per replication cycle. This high mutation rate causes significant variability in the surface proteins, especially hemagglutinin [70]. Increased variability inevitably leads to novel viral subtypes which are able to escape immune detection. This process is called antigenic drift. This unusually high rate of mutation is due to vRNA polymerase's lack of proofreading mechanisms [82]. Animals also almost certainly act as reservoirs for influenza [96]. Notably, animals with both α(2,3) and α(2,6) linkages may act as a type of melting pot that allows virus to undergo re-assortment. Re-assortment is the exchange of RNA segments between genetically unique viruses [96]. For example, the 2009 Swine Flu pandemic was likely caused by a single amino acid substitution in the protein PB1-F2 [19]. Rapid recombination and re-assortment can result in an even more rapid phenomenon called antigenic shift. Antigenic shift gives influenza its ability to quickly jump between species and create novel strains [35].

2.2.2 Influenza in Animals

Influenza is able to rapidly mutate and form novel subtypes. This allows seasonal epidemics in humans despite advances in vaccines and antiviral medications. However, influenza infects other animals as well. Notably, waterfowl and other aquatic birds are widely affected [33]. Swine are also prone to infection. Horses and some other animals may be infected as well but these infections are generally not transmissible to humans [96]. On the other hand, birds and pigs often transmit novel influenza strains to humans and act as long term reservoirs for subtypes [33, 96].

Aquatic birds and waterfowl acting as regular reservoirs has serious implications for human health. Domestic birds may also carry influenza. In many cases, all of these birds are asymptomatic. Generally, the receptors for bird influenza strains exist in the intestine. The virus may then be transmitted by fecal matter and in water. Ducks, for example, shed the virus heavily. A large percentage of the populations of waterfowl, especially juvenile birds, may be infected year round [97, 33, 96]. The permanent reservoir which birds provide influenza allows inter-seasonal mutations and the reemergence of viruses that had otherwise disappeared in humans.

Pigs may provide a similar environment for influenza virus. Pigs, uniquely, have both avian and human receptors. As such, they are able to act as an effective bridge between avian and human influenza [85]. In order for influenza to leap from birds to humans, mutations must occur on the binding sites of hemagglutinin. However, since swine can be infected with both human and avian strains, rapid antigenic shifts can occur [96]. This can promote the formation of novel combinations of surface proteins. This, by extension, can lead to pandemic strains of influenza since no prior immunity exists in human populations.

Aside from birds and pigs other animals may become infected with influenza. Generally they are not of concern to human populations. However, animals provide useful models for studying influenza. Many of the studies examining influenza transmission, vaccines and pathogenesis use guinea pigs or ferrets as model organisms [11].

2.3 Pathology

The pathology of influenza is well documented. The disease is an upper respiratory infection that lasts several weeks and is not usually life threatening. Most cases are clinically diagnosed, meaning they are made in the absence of laboratory testing. A combination of the time course, context and symptoms are used by physicians in outpatient settings to estimate influenza infections [61]. The CDC and WHO suggest using masks, washing hands and regularly disinfecting surfaces [6, 5]. Non-pharmaceutical interventions are also attractive and widely used since they cost less than antivirals and take less time to distribute than targeted vaccines. However, their efficacy is debated and not enough is known about the mechanisms of influenza transmission to make absolute recommendations [43, 13].

2.3.1 Mechanisms of Transmission and Infection

The mechanism of influenza transmission between individuals is of great importance in an epidemic or pandemic. The question of how best to prevent spreading of infection is difficult to answer. There are three main ways by which influenza virus may be transmitted. One, direct contact between an infected individual and a non-infected individual. This may be through shaking hands or other direct touching. Secondary contact via some surface such as a door knob may also play a role [47]. Two, large droplets may carry influenza virus. These droplets can be expelled by an infected person while coughing or talking. However, these large droplets generally fall out of the air after about 1m [47, 43]. Three, small aerosol droplets, generally defined as $< 5\mu$m, may be expelled by infected patients [16, 48, 47]. These small particles are likely the primary source of infection as they remain airborne for the longest time and are able to reach the lower respiratory tract [47].

Large droplets and aerosols are not distinct categories, however. Classification is based on size and droplets exist across a spectrum with distinctions between the categories blurring towards the center. In fact, as liquid evaporates, large droplets may become smaller aerosol particles mid-air. A significant number of infected patients expel viable amounts of viral material when coughing or speaking [52].These droplets can also be propelled across the room [48]. Mere proximity to an infected individual does not guarantee transmission though. Both aerosol and large droplets are important in transmission of virus, but they each face challenges entering the body.

Airborne virus, whether in large droplets or small aerosol particles must reach vulnerable tissue. Large droplets may contain more virus and that virus may be better protected from the environment. However, large droplets generally do not pass the upper respiratory tract. There, thicker mucus necessitates large amounts of viable viral material to infect cells [47]. Aerosols, on the other hand, may not contain as much viral material or be able to survive as long. They can infect lower respiratory epithelium, though. Additionally, they may be able to travel several meters away and remain suspended in the air for much longer [43, 47, 48, 92].

Once viable viral material reaches susceptible tissues in the respiratory tract, it is able to replicate [103]. From there, the virus spreads and causes the mild to moderate upper respiratory symptoms influenza is known for. In order to then infect a new host, the virus must be transmitted in one of the ways discussed. An infected patient sheds viral material for approximately 3 days. This shedding may begin prior to onset of symptoms. Additionally, younger children shed significantly longer than older children and adults [67].

2.3.2 Clinical Signs and Symptoms

Once a patient is infected with influenza, some time may pass prior to displaying symptoms [67, 103]. This is an important consideration when observing transmission and incidence rates, as a patient may spread the virus prior to displaying clinical symptoms and may delay seeing a doctor for several days after infection. Delayed reporting can create a lag in observed incidence levels.

After the appearance of symptoms, the illness can last for several weeks. Generally they are mild and not life threatening [5]. However, a large portion of hospitalizations and deaths occur as a result of co-infection with bacterial illness such as pneumonia [24]. In fact, many of the deaths during the 1918 Spanish Flu were caused by bacterial co-infections [62]. Most cases of influenza do not require hospitalization and are clinically diagnosed [5, 61]. Symptoms are variable and are generally similar to other upper respiratory infections (Table 2.1). Thus, accurate clinical diagnosis has an direct impact on ILI rates.

2.4 Immune Response

The respiratory tract is the main point of entrance for influenza viruses. The virus is able to infect upper and lower respiratory tissue, initiating innate and adaptive immune responses. Dendritic cells are the primary innate immune mediator, recognizing viral particles [12]. The adaptive immune response is handled by effector T cells responding to viral antigens. These cells are responsible for balancing adaptive responses as well as regulating the inflammatory response [12]. Both innate and adaptive immune responses are important for viral clearance and provide immunity to re-infection.

2.4.1 Innate Immunity

The first line of defense to influenza, and most other diseases, is the innate immune system. Given that influenza is a respiratory illness in humans, our focus will be on

Table 2.1 Common Influenza Symptoms and Their Clinical Frequency. [61, 6]

	Symptoms	*Clinical Frequency*	*Description*
(i)	weakness	94%	generalized fatigue and lack of strength
(ii)	myalgia	94%	muscle pain or soreness
(iii)	cough	93%	usually a hacking, dry cough
(iv)	nasal congestion	91%	“runny nose” and clogged sinuses
(v)	subjective fever	90%	feeling feverish without a measurement of temperature
(vi)	objective fever	68%	measured temperature of above 38 °C
(vii)	loss of appetite	92%	reduced desire to eat
(viii)	headache	91%	generalized pain to any region of the head

the barriers and responses present in the upper and lower respiratory tracts primarily. Before infection can begin, influenza must travel into the body through the mouth or nose. From there, it can move to the upper or lower respiratory tract and infect epithelial cells. These cells are not defenseless though. The first barrier to infection is the layer of mucus that constantly coats the respiratory tract. This mucus is rapidly cleaned and replaced. It captures and flushes out influenza, along with other invaders, sending them down the esophagus and into the stomach. Here strong stomach acid destroys most bacteria and viruses [18]. However, as noted earlier, neuraminidase may be responsible for helping viruses avoid becoming trapped in mucus layers.

Once cells are actually infected, vRNA may be recognized by pattern recognition receptors or type 1 interferons which may promote cytokines and IFN-stimulated genes [39]. Cytokines promote a variety of systematic and local immune responses, including promoting inflammation. Additionally, cytokines recruit non-specific immune cells such as macrophages and phagocytes to clear infected cells [39, 34]. IFN-stimulated genes produce a variety of proteins that aid in defense [39]. One example is Myxovirus resistance protein 1 (MxA) which is the product of MX1 and may help prevent nuclear import of viral components [36]. Another example is Interferon-induced transmembrane protein 3 (IFTM3) which inhibits viral release, preventing new viruses from infecting more cells [14]. Finally, Tripartite motif-containing 22 (TRIM22) protein targets viral nucleocapsids for degredation [22].

While the innate immune response plays an important role in preventing and eventually eliminating infectious agents, it may have negative effects as well. Influenza is noted to cause a variety of systematic symptoms [61]. However, influenza is not a systematic illness. In extreme cases the immune response to viral infection can be harmful. And in some cases, increased host response to infection can lead to increased disease severity and mortality [72].

2.4.2 Adaptive Immunity

While the innate immune system hinders the influenza infection, the adaptive immunity is responsible for clearing the body and preventing reinfection. There are two primary adaptive immune responses, the humoral and the cell response. Both play an important role in fighting viral infection.

The humoral response to influenza is mediated by B-cells. These B-cells produce antibodies that primarily target hemagglutinin and neuraminidase [45]. Anti-HA antibodies bind to the active site of HA and prevent attachment to sialic acid. This results in inhibition of viral attachment to host cells. Anti-neuraminidase antibodies limit viral spread by preventing sialic acid cleavage by neuraminidase. There are also M2 specific antibodies that have been shown to effect virulence [45].

The primary classes of anitbodies responsible for anti-influenza activity are IgA, IgM and IgG. These are common mucosal and serum antibodies [91]. These antibodies, and the B-cells responsible for producing them, are what allows seasonal influenza vaccines to be effective. The cell mediated response relies on CD4, CD8 and regulatory T-cells [91, 45]. Notably, cytotoxic T-lymphocytes eliminate infected cells [12].

2.5 Epidemiology

Seasonal influenza is a constant concern despite modern advances in vaccines and antiviral medications. Influenza tends to spread rapidly and it is especially potent during the winter when people tend to be in close contact indoors. So, unsurprisingly, seasonal epidemics occur in the late fall through early spring [5]. Patterns form year to year and are of importance to healthcare professionals trying to develop vaccines and prepare for potential pandemics.

2.5.1 Seasonal Influenza

Seasonal influenza creates a regular, repeating pattern from year to year. Influenza cases begin to rise in October, peak in January or February, and trail off into April. A close-up of the 2013-2014 influenza season can be seen in Figure 2.4. The regular nature of influenza epidemics is only surface deep, though. While the general pattern remains similar year-to-year, there exists substantial variation. This variation can be clearly seen in Figure 2.5. Years differ in overall incidence, peak, onset, and end. Notably, pandemic years such as the 2009 Swine Flu may be significantly different from the norm. Pandemic years in general are characterized by increased virulence, disproportionate effects on the young and elderly, and summer illnesses [62]. Attempts to model the yearly variation have been met with varying success.

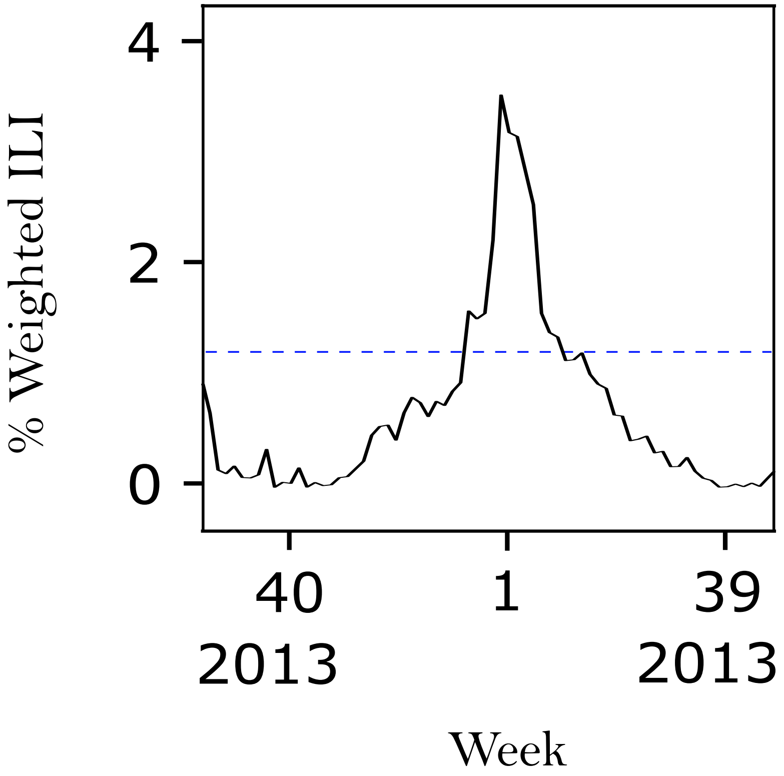

Figure 2.4 Graph of 2013-2014 influenza season ILI reveals winter peak. 2013-2014 was a typical flu season. ILI incidence rises slowly through October and November, tipping over the regional baseline at week 50 of 2014. The season peaks shortly after and the dips below the baseline again in April at week 17 of 2015.

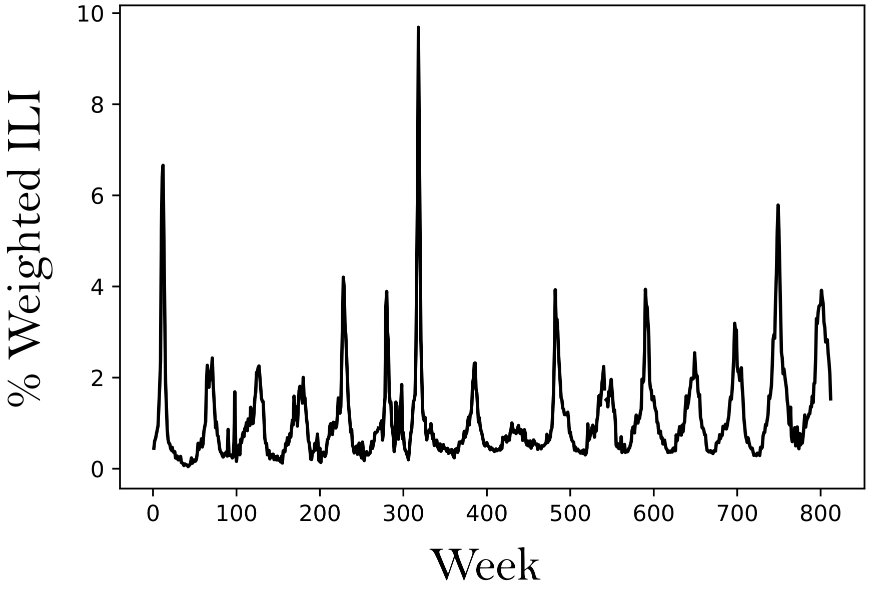

Figure 2.5 Weekly ILI from 2003 to 2018 reveals regular, repeating outbreaks. Peaks influenza incidence occurs each year during winter months. The exception is the 2009 flu season, now known as the Swine Flu pandemic, which can be found centered at approximately week 300. This pandemic season was unusual in that ILI incidence remained elevated through the spring and summer.

There is no firm consensus on what causes the seasonal variability, but temperature, dry air, and host immune irregularities may play a role [89, 25, 54]. People generally spend more time indoors during colder weather and are therefore in regular close contact. School closures, for example, are correlated with reduced ILI incidence, suggesting that close proximity of infected and susceptible individuals is a driver for seasonal spikes in influenza incidence [40]. Additionally, despite prior exposure, novel viruses emerge that can evade host immune responses. This further increases yearly variability [54]. Ultimately, yearly variability may be due to very small changes in a multitude of variables that are amplified by population dynamics [23]. Interestingly, tropical regions do not show strong seasonality. Instead they have generally flat ILI incidence that varies with rainy season [89, 90].

2.5.2 Pandemics

While seasonal influenza epidemics are relatively predictable in many aspects, the threat of novel sub-types emerging is constant. Influenza's rapid mutation rate, combined with persistent infection in reservoir species such as waterfowl and swine, allows it to outpace host immune response. Influenza pandemics caused by novel viruses occur irregularly and are difficult to predict. Early prediction is critical to preparation and vaccine development, though. A rapid response can drastically reduce disease burden and prevent excessive mortality.

Most notable of these pandemics was the 1918 Spanish Flu. The exact origin of the H1N1 influenza that caused this pandemic is not known, but it likely moved from swine to humans [42]. This was an especially virulent disease that disproportionately affected young, healthy individuals [62]. Its virulence was likely caused by a single mutation of PB1-F2 [19]. Over the course of the pandemic, the Spanish Flu was responsible for an estimated 40-50 million deaths worldwide. Most of these deaths were probably a result of bacterial co-infection, though [30].

Other pandemics of the 20th century include the 1957 H2N2 Asian Flu and the 1968 Hong Kong Flu [42]. Most recently, the 2009 pandemic emerged as a result of a novel H1N1 virus. It was notable for increased virulence and disproportionate effects on the elderly and children. The disease was generally mild but there was potential for serious complications [73, 29]. A targeted vaccine was rushed out in record time during the 2009 pandemic. The vaccine had a significant impact [31]. Additionally, vaccines that contained the same H1N1 continued to be effective several years after the pandemic [27]. These facts reinforce the need for early predictions of pandemics and emphasize the effect a timely response can have.

2.5.3 Vaccines

Vaccination is a critical preventative measure that is useful during regular seasonal flu seasons as well as during pandemics. Because of antigenic drift in the influenza virus, new vaccines must be developed yearly [71]. These vaccines have effectiveness rates that range from 10% to 60% [10, 93]. Vaccine effectiveness is extremely variable and illustrates the difficulty in predicting relevant strains (Table 2.2).

Targeted vaccines were effective during the 2009 Swine Flu pandemic in the United states [31]. However, strain-specific vaccines are not a practical defense against pandemics [71]. Generally pandemics are sudden and unexpected. Thus, targeted vaccines need to be developed rapidly and may be released significantly after a pandemic begins.

2.5.4 Climate and Other Driving Factors

Seasonal variations in influenza are difficult to predict. The driving factors for yearly differences in incidence rates are not well understood [89]. No true consensus exists on the effect of individual climate variables, or other non-climate variables [80]. However, given years of research, there are several strong candidates for seasonal drivers of influenza.

Table 2.2 Vaccination Rate and Effectiveness Varies Considerably From 2003-2018. Vaccination rates calculated from data available on CDC FluVaxView [1].

Influenza Season	*Vaccine Effectiveness (%)*	*Vaccination Rate (%)*
2017-2018	38 [76]	55.0
2016-2017	40 [26]	41.7
2015-2016	48 [41]	46.8
2014-2015	19 [106]	45.6
2013-2014	52 [27]	47.1
2012-2013	49 [60]	43.7
2011-2012	47 [68]	42.1
2010-2011	60 [93]	39.2
2009-2010	56 [31]	38.8
2008-2009	41 [1]	34.2
2007-2008	37 [9]	31.7
2006-2007	52 [10]	28.7
2005-2006	21 [10]	25.6
2004-2005	10 [10]	20.0
2003-2004	52 [2]	12.7

The most strongly correlated climate variable in temperate regions is temperature [56]. Lower temperatures in particular may drive behavior that can increase flu transmission, such as close contact and recirculated air [56]. However, temperature itself may contribute to increased viral survivability and prolonged shedding [55].

Humidity, which also may affect viral survivability, shows up as a consistently good predictor in a variety of scenarios in both temperate and tropical regions [21, 81, 90]. Humidity combined with temperature may both be strong drivers as close contact during winter, and other non-climate explanations, may not be sufficient. Close contact during summer months does not produce influenza epidemics in temperate regions [20]. Secondary to humidity is precipitation, which is a strong predictor in tropical regions [21].

Additionally, low UV index may contribute to seasonal trends [38]. This may be in large part due to the role of vitamin D in the immune system. Vitamin D is produced via sun exposure and a lack of vitamin D may result in immune deficiencies. UV index alone is a strong enough predictor to explain many variations in seasonal epidemics [15].

Overall there are a large number of factors, more than have been covered in this section, that contribute to seasonal epidemics. Variation from year to year is difficult to predict as no consensus on the underlying mechanisms of seasonal influenza exist. In all likelihood, variation in seasonal trends may be the result on minute changes in one of many factors. Given previous research, temperature, humidity and UV index provide the best predictors to influenza rates. However, the underlying mechanisms that create these correlations are not understood at this time.

2.6 Modeling

Influenza is a seasonal disease and almost guaranteed to regularly affect millions. Providing hospitals and public health professions ample time to prepare is critical to mitigating the impact of flu seasons. Given the complex nature of the variables involved in yearly variation, and the poor understanding of the underlying mechanisms, a wide variety of models have been produced. These models seek to forecast everything from broad yearly trends to granular incidence data.

Notably, each year the CDC holds a competition to forecast influenza seasons. The challenge involves predicting 4 weeks out from reported flu incidence, predicting peak week, peak intensity and onset week. A variety of models have been submitted. Models include classic mechanistic models based on SIR (susceptible, infected, recovered) as well as statistical models based on machine learning. Many models use a combined approach for predicting various portions of the challenge. In general, these models predict one week better than the CDC historical average, but the predictive effectiveness falls off towards four weeks [74].

2.6.1 Modeling Approaches

Two basic approaches exist for modeling diseases. A mechanistic approach seeks to break disease transmission into discrete segments that can be manipulated and combined to produce accurate reproductions of disease dynamics. In general, mechanistic models are based on biological principles and explain underlying mechanisms [74]. Some examples include a 2013 study that models flu trends in Israel. The model is a modified SIRS (susceptible, infected, recovered, susceptible) model and basic climate and viral evolution factors. It successfully models general trends [8]. Another model making use of a modified SIRS model was able to predict more granular data out to 7 weeks [80]. Mechanistic models such as these can provide insight into the mechanisms that drive a biological process. However, they struggle to model phenomenon that are not well understood. As mentioned in the previous section, there is no consensus on the driving factors affecting seasonal flu.

The alternative approach, statistical modeling, is better able to handle underlying uncertainty. Statistical approaches to modeling are based on observations and raw data. Predictions are made based on statistical trends. In many cases, machine learning is applied. This approach can allow the model to detect trends not readily understood or view-able by the designers of the model. Unfortunately, statistical models provide no explanation of the underlying mechanisms [74]. A variety of techniques exist in this category. ARIMA (Auto-Regressive Moving Average) may be applied to flu data, as was in a 2010 study that explored climate variables and time-lag effects [87]. SARIMA, or seasonal ARIMA, was also a common technique in the CDC challenge [74]. Machine learning can also be readily applied for predictions. A 2019 study used Random Forest modeling to evaluate climate factors including UV index [38].

2.6.2 LSTM and Neural Networks

One machine learning approach that has gained popularity in recent years is LSTM or Long-Short-Term-Memory. This technique, when applied to influenza, performed better than random forest regression, support vector machines and ARIMA [105]. LSTM based neural networks have been used to assess social media data for flu prediction [95]. LSTM based neural network models are only beginning to be applied to influenza trend prediction [53].

LSTM based neural networks provide some distinct advantages over other types of neural networks, which themselves provide advantages over other predictive techniques. Neural networks are structured as an interconnected network of nodes called neurons. These neurons represent self-contained sets of algorithms that output values based on their input. Neural networks allow models to learn vast amounts of data and detect patterns that would be otherwise impossible to extract. Two main types of neural networks exist, feed-forward and recurrent. In feed-forward networks, the output of the previous node is feed into the next node. In recurrent networks, time series data may be used, as results are fed back to previous nodes [44].

LSTM nodes were designed to allow for time-series forecasting. Specifically, they seek to solve the problem of disappearing or exploding gradients that is common in recurrent neural networks [28]. Gradients are an integral part of neural networks, they affect the ”on/off” signals of the individual nodes. Depending on the data set and hyper-parameters of the model, gradients can produce NA values. Essentially they run out of bounds. LSTM nodes circumvent this problem by introducing a CEC or constant error carousel [37]. The CEC allows for gradients to remain unchanged from one node to the next. The more recent addition of a ”forget gate” allows the LSTM node to reset, further reducing gradient runaway [28]. The basic structure of an LSTM as implemented in Keras includes a forget gate, and input gate and an output gate (2.6). LSTM based neural networks allow for complex time-series forecasts. They are an ideal candidate for influenza prediction and provide a relatively novel foundation for forecasting.

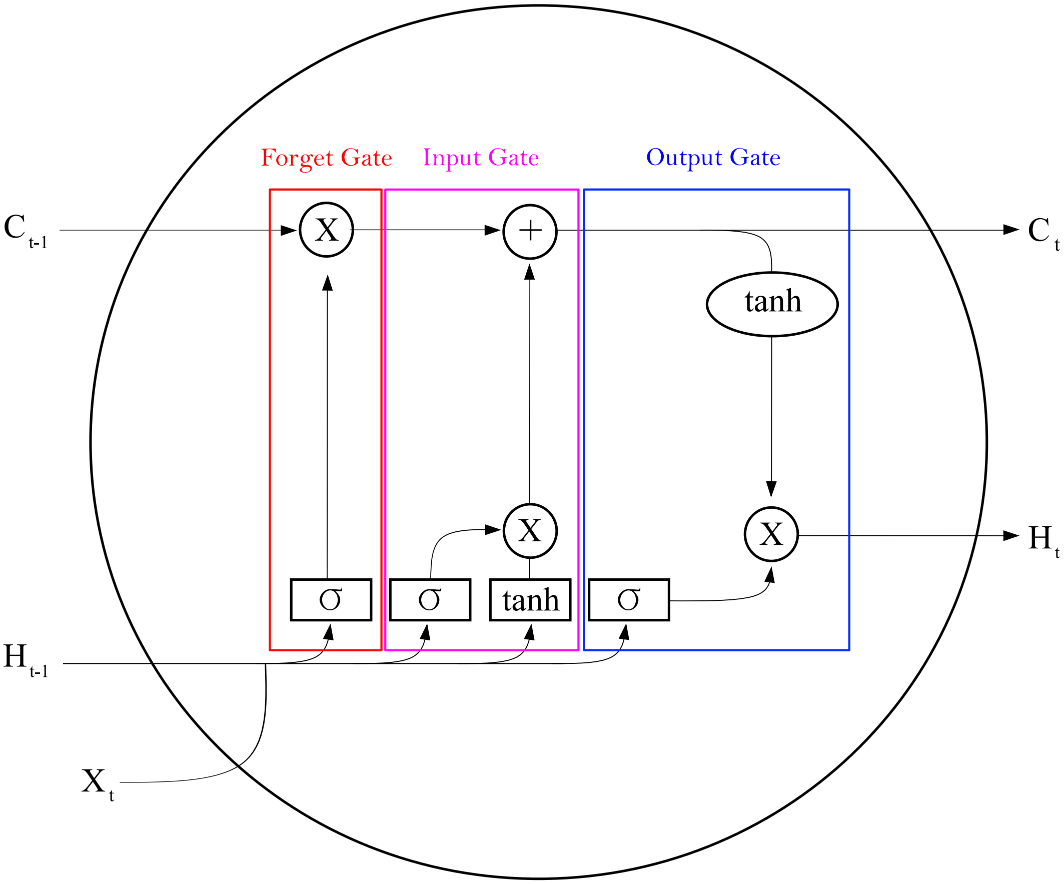

Figure 2.6 LSTM solves the problem of disappearing gradients by implementing a CEC and a forget gate. LSTM nodes contain a forget gate, an input gate and an output gate. This architecture seeks to mitigate the effects of disappearing or exploding gradients. σ denotes a hard sigmoid function, *tanh* denotes a hyperbolic tangent function. X and $+$ denote a multiplication and addition process, respectively. $C_{\text{t-1}}$ is the memory from the previous LSTM node. $H_{\text{t-1}}$ is the output from the previous node. X_{t} is the input.

CHAPTER 3

MATERIALS AND METHODS

In this following two sections, the data acquisition process and the model building process are detailed. Supplemental information on the data sets used is available in the appendix. Additionally, select code segments are available for review. All code related to the models may be found online. Data was processed in MiniTab and R. Final data manipulation was done using Python. The models were designed and constructed in Python using TensorFlow 2.0 BETA and the Keras API. TensorFlow used GPU acceleration. The computer specs used to run the models are: Intel i7-3829 @ 3.60GHz, 64GB DDR3 RAM, RTX 2080 Ti 11GB.

3.1 Data Compilation

A source of data on influenza trends was identified. The Center for Disease Control collects data from public health labs and private doctors offices. This provided the most consistent data spanning 2 decades. The CDC posts weekly ILI rates and has weekly records from 1998. This data is provided on a national level, in some cases a state level and a regional level. In order to narrow the focus of this paper, the CDC region 1 New England was chosen. It has distinct seasons, relatively uniform climate, it is geographically continuous and climate data is readily available. Additionally, it has continuous data from at least the 2003-2004 season. New England, for the purposes of this paper, contains Maine, Connecticut, Rhode Island, New Hampshire, Vermont and Massachusetts.

Once New England was selected, the initial data set from the CDC Flu View was downloaded as a Comma-Separated-Values file. This set contained ILI percents, total patients, and information on sub-typing. More information on this data, and a sample can be found in the appendix. The data was imported into R. Data ranged from week 40 of the 2003-2004 flu season to the current 2018-2019 week. The data was trimmed to include up to week 21 of the 2017-2018 season.

In addition to the raw data, the CDC calculates a regional base line for each year. This baseline is calculated by taking the mean percent ILI for non-influenza weeks from the three preceding seasons and adding two standard deviations. The CDC defines a non-influenza week as periods of two or more consecutive weeks in which each week accounted for less than 2% of the seasons total number of specimens that tested positive for influenza in public health laboratories. This data was available from the 2007-2008 season onward. In order to fill in missing baselines for the previous several seasons (from the 2003-2004 season to the 2006-2007 season) the CDC procedure was followed as close as possible. Beginning with the 2003-2004 season, a 1 year baseline was calculated since years prior to 2003 did not report off season ILI levels. Then the next year had a 2 year baseline and so on until a full three year baseline was available. The estimated baselines were adequate for the purposes of this model. All data was weekly. A total of 816 weeks were used.

After acquiring flu data parameters were chosen. A set of climate, population, and epidemiological factors had to be identified. An informal survey of recent papers addressing the effect of climate on seasonal influenza trends was conducted. From this, several promising climate variables were compiled. Temperature, humidity and UV index were best correlated with influenza trends and were supported by a multitude of studies. A 2016 study on influenza trends in the tropics further supported a link between humidity and influenza outbreaks. El Nino years also showed higher than usual influenza activity. A 2013 study conducted on data from Israel indicated that along with climate, incorporating antigenic drift and immunity loss increased accuracy of multi annual influence forecasting. From the available literature a list of potential climate variables was compiled.

In addition to climate factors, population density, travel patterns, and time spent indoors were identified as potential drivers. Heating day and cooling day count was used to represent time spent indoors. Heating days specifically is defined as any day below a set temperature, usually 65 C. This is an industry measure to estimate heating costs, but might also provide a proxy for cold days when people are more likely to spend time indoors. Once these initial parameters were identified, dthe climate data was accessed.

Climate data was taken from the National Oceanic and Atmospheric Administrations Climate Data Online. In order to provide a sample representative of the region, a single monitoring station was selected from each state for a total of 6 weather stations. These stations include Hartford Bradley Airport, Connecticut; Boston, Massachusetts; Augusta Airport, Maine; Mt. Washington, New Hampshire; Providence, Rhode Island; Montpelier, Vermont. Most data was available as daily averages. Some data was only available as monthly averages. All data was converted to weekly data and trimmed to match the CDC data already collected. The mean of each weeks data was then calculated to produce regional weekly data, which was included in the final data set. The climate factors used were: average temperature, average wind speed and precipitation.

In addition to the climate data, population data was taken from the U.S. Census Bureau. This data included population totals and immigration data. This data was added to the master sheet.

Finally, vaccination data was taken from the CDC. Vaccination rates were collected on a regional bases from the CDC website. Estimated vaccine effectiveness was extracted from the CDC website as well as scientific papers. See the appendix for more details.

The final data set was limited to data that was available regionally on a consistent basis. The only major missing climate data that was estimated was wind speed from Rhode Island. Missing data was filled in with average data from the previous 10 years. The tail of the actual data set used is available to view in the appendix. Due to availability constraints some promising parameters were not included. These parameters were UV index, absolute humidity, El Nino, population density and regional travel.

3.2 Building the Model

The master data set was then prepped for input into a model. Data originally organized in a simple dataframe. However, in order to allow use of various node types, the data was reshaped into a 3 dimensional array. The data was ultimately broken into time-steps that represented 1 weeks data. Prior to reshaping, data was standardized using the following equation:

$$\frac{(x - mean)}{standard\ deviation}$$

Once the data was reshaped and standardized, it was broken into training and testing sets. In order to make the best use of limited data, several configurations were used. The data was split into a variety of segments that were then used to train and test the model in order to ensure generalizability. The data splits can be seen in Figures 3.1, 3.2 and 3.3.

Because of the time factor the data was not shuffled when training and testing. However, separate models were trained and tested on shuffled data to determine a random baseline and act as comparison. Select code segments can be found in the appendix.

Once the data had been prepped, the actual model design began. In order to develop a baseline for comparison, a simple deep neural network was constructed. All specific model architectures used are explained in detail in the appendix. This initial model predicted weekly ILI. Initially, the model was fed only influenza data.

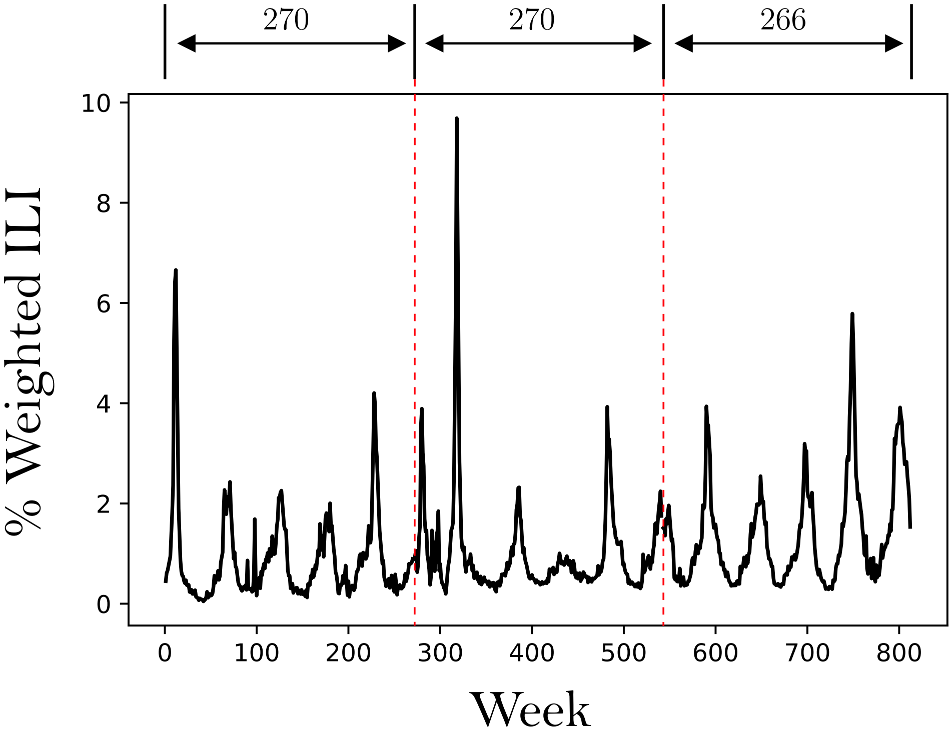

Figure 3.1 Data split into 270 week segments. The data was split into 2 sets of 270 weeks and one set of 266 weeks. The model was trained on 2 sets and tested on 1, with the training and testing sets rotated for each trial.

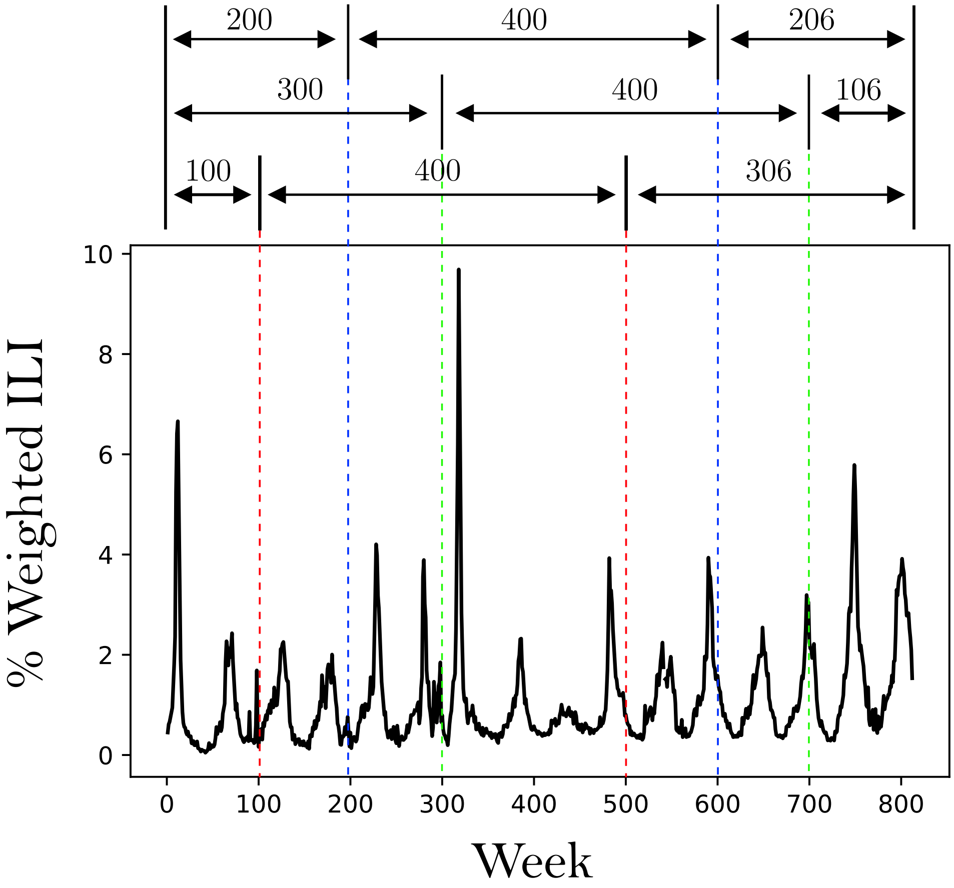

Figure 3.2 Data split into 400 week training sets. The data was divided to allow a 400 week training set. The training set was then shifted to evaluate generalizability.

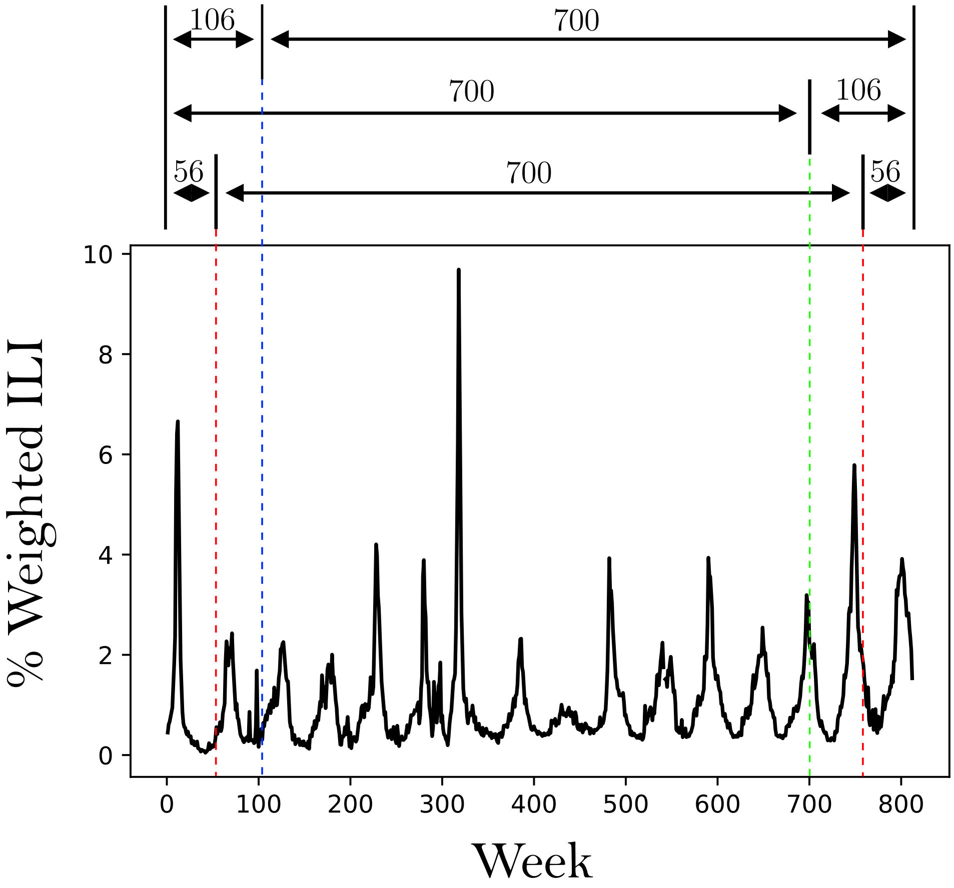

Figure 3.3 Data split 700 week training sets. The data was divided to allow a 700 week training set. The training set was then shifted to evaluate generalizability.

Other parameters were added individually to asses impact on accuracy. Mean square error and mean absolute error were used as metrics to determine model accuracy and control learning. In this first model, 3 dense layers using Rectifier linear unit (ReLU) algorithms were implemented. The ReLU formula is defined as:

$$y = max(0, x)$$

By continually adding parameters, the most effective predictors could be identified. A reduced data set was then compiled and a new model was designed.

The second iteration of the predictive model relied on time-series forecasting to predict ILI rates one or more weeks in advance. In order to test features of the model, only two layers were implemented to begin with. An input layer containing a long-short term memory (LSTM) layer with 4 nodes and an output dense layer with a single node was used. These nodes used the default hyperbolic tangent (tanh) activation and hard sigmoid recurrent activation, shown below:

$$Hyperbolic\ Tangent: \quad tanh(x) = \frac{sinh(x)}{cosh(x)}$$

$$Hard\ Sigmoid: \quad max(0, min(1, x * 0.2 + 0.5))$$

From there, nodes and layers were added incrementally until gains slowed. Initially, only 1 week was predicted. However, a function that looped predictions was written to allow recurrent predictions to be made. Thus from this point on, the model was able to predict further than 1 week into the future.

The final iteration of the model (3.5) was a recursive deep neural network made up of a bidirectional LSTM input layer, two bidirectional LSTM hidden layers (3.4) and a dense output layer with variable output nodes .

This model was tuned incrementally to achieve the best predictions. In order to achieve the best performance.

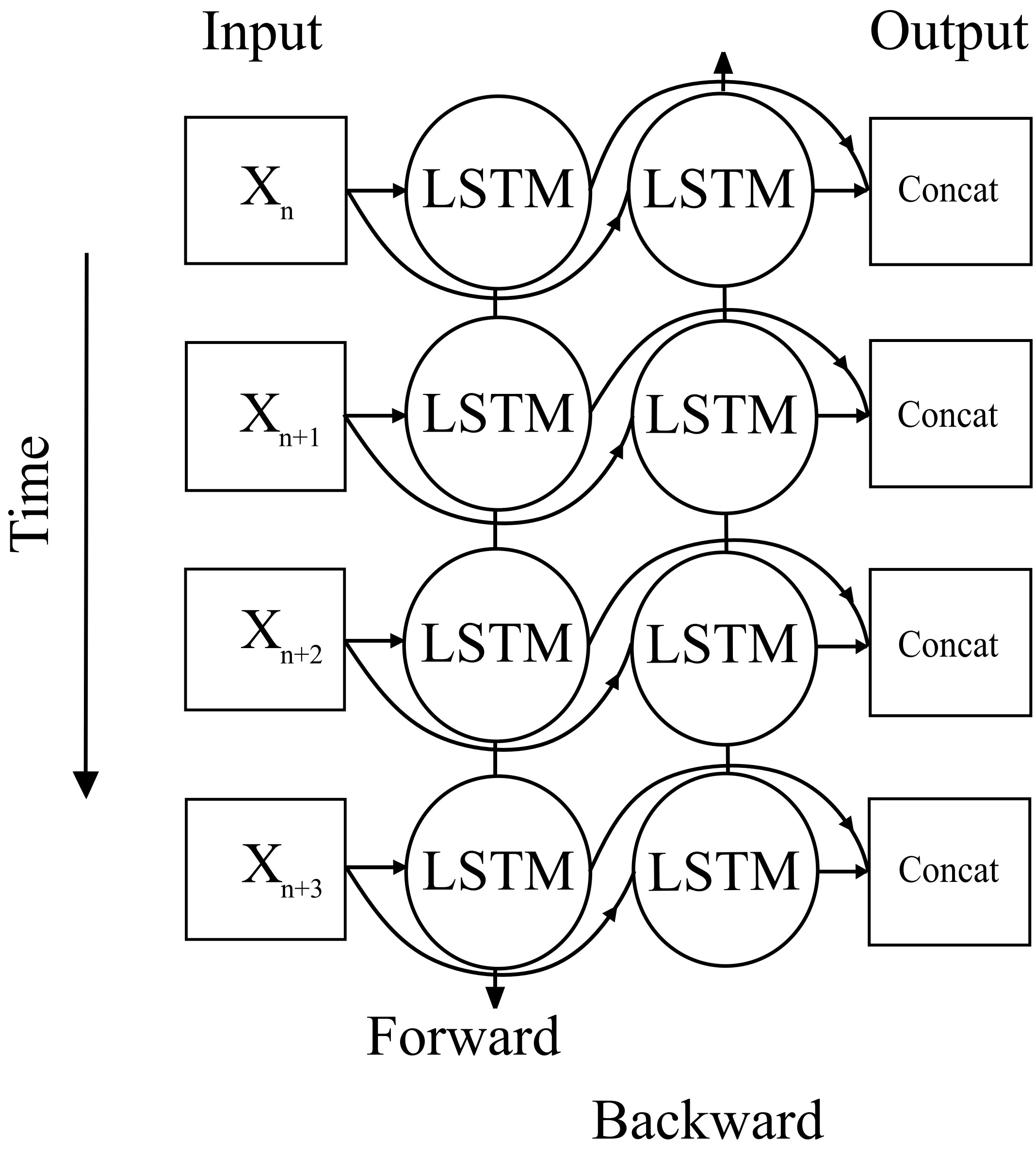

Figure 3.4 Bidirectional LSTM layer diagram.

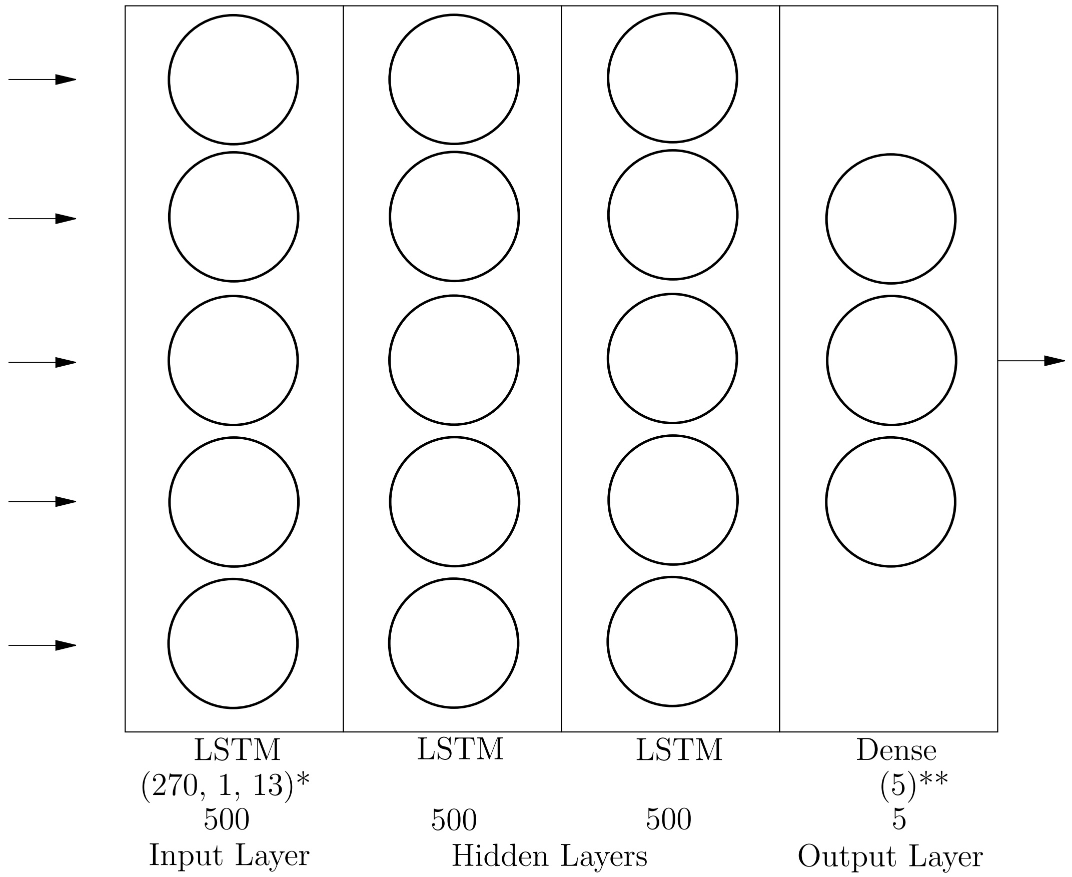

Figure 3.5 Final model architecture. The final model contained 4 layers total. An initial 500 node LSTM input layer with a variable shape, 2 hidden LSTM layers with 500 nodes each and a dense output layer with a variable output shape.

* the input shape varies with data shape

** output shape varies with label shape

Once satisfactory base performance was achieved, comparisons were made between training sets to confirm generalizability of the model, as well as identify any potential data leaks. Then, a comparison of two different predictive methods was made. One model predicted to out to 10 weeks automatically. The other model predicted out to 1 week and recursively predicted the next 10 weeks using the function mentioned early. Details of this function can be found in the Appendix. A comparison of various time lags ranging from *t - 1* to *t - 52* was made.

Individual climate and population variables were also evaluated. After evaluation, any variables found to negatively impact performance were removed from the data set. The resulting data set will be referred to as the *reduced data set* as opposed to the *full data set* that includes all data.

3.3 Predicting Outliers

Two models of identical architecture were trained on different data sets. The first model, called $\mathrm{P}_{\text{included}}$, was trained on the reduced data set containing all weeks from 0 - 540. The second model, $\mathrm{P}_{\text{removed}}$, was trained on the same time span, except outlier years were removed. Outlier years were chosen based on standardized percent ILI. Any year with a standardized percent ILI above 4% was removed from the training date. Both models were then tested on weeks 541 - 806. Divergence between predictions was then used to extract a signal indicating an outlier year.

CHAPTER 4

RESULTS

Predictions were made for tests sets of various lengths and frames. Baseline performance was determined, then the most effective time lag was selected, and finally, the data set was evaluated. Overall performance for each model was established using absolute mean error (MAE), mean error, standard deviation, and visual analysis. Week 1 predictions were the most accurate. Predicting further than 5 weeks was influenced heavily by time lag, modeling method, and data selection. The most significant increases in performance were achieved by tuning the time lag and by using the recursive prediction function. The model was applied to outlier prediction by combining the outputs of two identical models trained on differing data sets. This multi-model approach was able to detect a weak signal preceding an outlier year.

4.1 Determining Baseline Performance

Once a functional model was created, baseline performance and generalizability were evaluated. Using the complete data set and a time lag of one week, nine different training sets were used to train models. These training sets were divided into three groups of 400 week training sets, 540 week training sets and 700 week training sets. MAE was used to determine relative performance along with visual interpretation of predictions. MAE was recorded for weeks 1, 5 and 10 (4.1). The best performance was achieved when predicting one week in advance. Both MAE and the standard deviation of the error rose substantially by week 10. Two sample t-tests were used to determine significant differences between week 1 predictions from each training set. There was significant difference between different frame shifts within all three training-set-length groups. The mean increase in MAE from week 1 to week 10 was 0.412. There was no significant difference between the 540 and 700 week training sets, although the 400 week training set performed significantly worse. Moving forward, 540 week training sets were used for testing as they provided sufficient predictive ability and were easier to manipulate.

In general, model performance degraded as prediction week increased. Additionally, the model consistently under-predicted values at weeks 5 and 10 (4.1). As true percent ILI values tended towards high values, the model also consistently under-predicted. The presence of extreme outliers in the test set, notably the 2009 Swine Flu pandemic, reduced predictive performance and resulted in under-prediction.

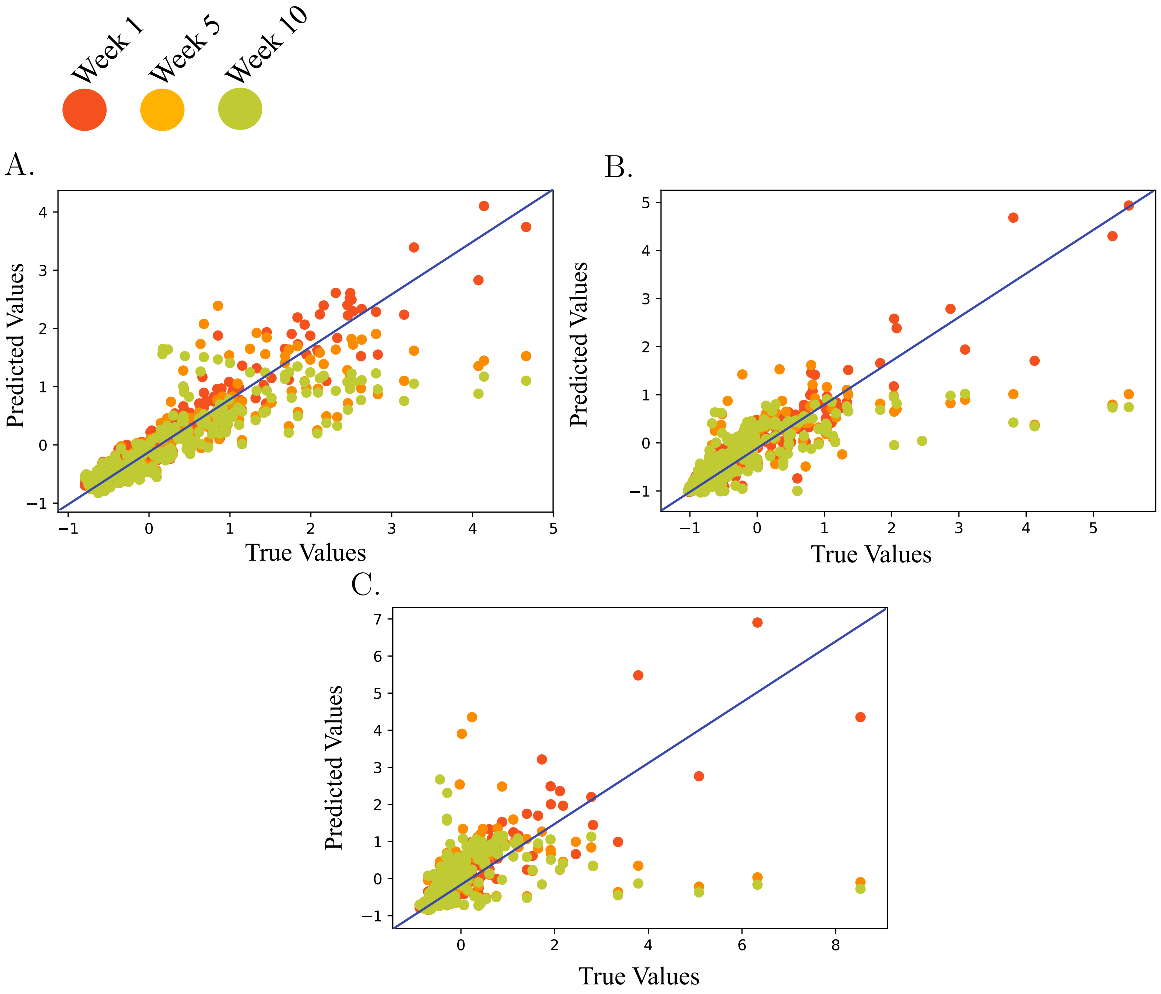

Figure 4.1 Prediction accuracy degrades as prediction time increases as well as in the presence of outliers. 1 week predictions have the highest accuracy. When predicting out to 5 and 10 weeks, predictions are worse, especially at extreme values. Predicting an set containing the major outlier year, 2009 Swine Flu, accuracy degrades (**C**). The recursive model with a 4 week time lag was used. Three versions were trained on weeks 0–540, 270–806 and 0–270 and 540–806 weeks. **A**, **B** and **C** depict predicted vs true percent ILI for each training set, respectively. Panel **C** shows predictions from the test set containing the 2009 Swine Flu pandemic. All values are percent weighted ILI.

4.2 Evaluating Variables

To fine tune the model and obtain the best predictions possible, climate and population variables were evaluated. Prior to this, the complete data set was used to make predictions. In order to determine how the variables may be effecting

Table 4.1 Mean Absolute Error of Various Training Sets Is Significantly Different Within and Between Training Sets. 400 week training sets provided the worst performance. There was no significant difference between using 540 week and 700 week training sets. There was significant differences between training sets that included the 2009 pandemic and those that did not.

Training Set		*Prediction Error (MAE)*		
		Week 1	*Week 5*	*Week 10*
400 Weeks	Weeks 100 - 500	0.6630	0.6637	0.6664
	Weeks 200 - 600	0.5370	0.5274	0.5394
	Weeks 300 - 700	0.3374	0.6682	0.6440
		0.5124	**0.6197**	**0.6166**
540 Weeks	Weeks 0 - 540	0.3103	0.4678	0.5792
	Weeks 0 - 270 & 540 - 806	0.3860	0.5399	0.5563
	Weeks 270 - 806	0.3130	0.3609	0.3878
		0.3364	**0.4562**	**0.5077**
700 Weeks	Weeks 0 -700	0.5227	0.5806	0.7309
	Weeks 53 - 753	0.3086	0.6008	0.6094
	Weeks 106 - 806	0.4263	0.6601	0.5861
		0.4192	**0.6138**	**0.6421**

predictions, they were removed one by one. The data set became progressively smaller until only data columns 'percent ILI', 'Week' and 'Year' remained. Temperature was the most important variable for predicting Week 1. Precipitation also had a significant effect when removed. Removing either of these variables reduced performance of week 1 predictions. Removing population and vaccination data appears to have improved predictive power substantially. Removing temperature and monthly precipitation, weekly precipitation and weekly temperature decreased predictive performance. The best predictions were obtained with a data set containing only precipitation and average temperature (4.2).

Table 4.2 Temperature Is the Strongest Predictor for ILI. Parameters were removed one by one in descending order. Models were trained on weeks 0–540 and tested on weeks 541–806.

Parameter	*Prediction Error (MAE)*
Base Model	**0.213**
Average Wind Speed - Monthly	0.182
Precipitation - Monthly	0.218
Average Temperature - Monthly	0.204
Population	0.195
Vaccine Effectiveness	0.195
Vaccination Rate	0.187
Average Wind Speed - Weekly	0.163
Precipitation - Weekly	0.185
Average Temperature - Weekly	0.231

4.3 Comparing Standard and Recursive Predictions

Once baseline performance was established, standard predictions could be compared to recursive predictions. Predictions using the standard model and a time lag of one week was used as a baseline for evaluation. Time lags of 4, 12, 16 and 52 weeks using the standard model were compared to baseline. A time lag of 4 weeks provided an average decrease of 0.1400 percent ILI error across weeks 1, 5 and 10. The greatest improvement was seen in week 10 predictions. Week predictive performance degraded as the time lag increased past 4 weeks (4.3).

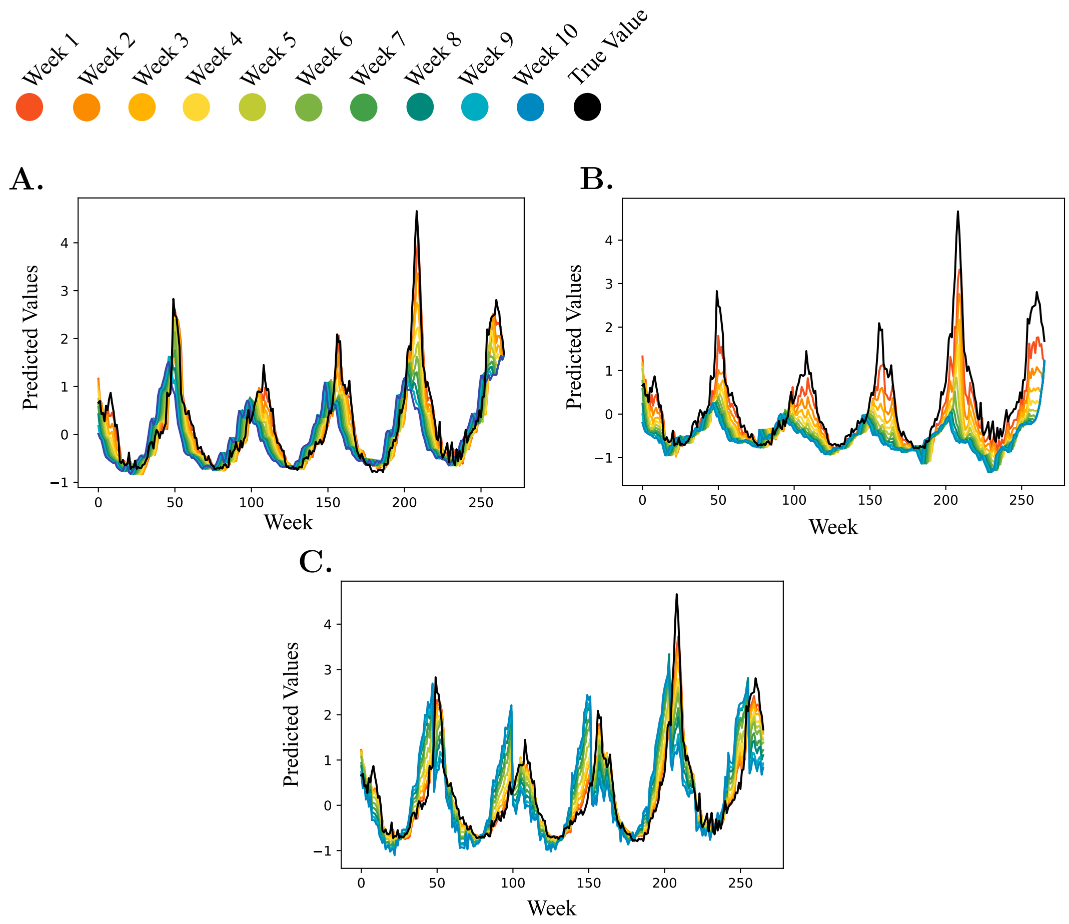

Figure 4.2 ILI predictions accuracy varies with time lag and prediction method. Using a recursive function with a 4 week time lag provided the most balanced predictions. Week 1 predictions compare well with label data. The predictive accuracy drops off as predictions range further out, however they remain better than alternative methods (**A**). Reducing time lag using recursive predictions causes under prediction across the entire prediction range (**B**). The baseline model using a 1 week time lag produces good 1 week predictions that rapidly degrade over the range of predictions (**C**). All predictions were made using a model trained on weeks 0–540 and tested on weeks 541–806. Predicted values are percent weighted ILI. Data plotted here has been standardized resulting in negative values on the y-axis. NOTE: Predicted values are offset due to plotting based on prediction start week, thus peaks for week 10 predictions appear slightly delayed.

Table 4.3 A 4 Week Time Lag Provides Best Predictive Performance. Increasing time lag degraded week 1 prediction performance using the standard model but improved week 5 prediction performance. Using a 4 week time lag increased overall performance of the recursive model vs. baseline

Time Lag	*Prediction Error (MAE)*		
	Week 1	*Week 5*	*Week 10*
Baseline	**0.3412**	**0.4485**	**0.4963**
t-4	0.2903	0.2882	0.2876
t-12	0.4000	0.4867	0.4709
t-16	0.3831	0.4592	0.4496
t-52	0.4838	0.5127	0.4868

4.4 Predicting Outlier Years

When outliers were removed from training data, predictive power was reduced in a regular way. The $\mathrm{P}_{\mathrm{removed}}$ model was unable to predict outlier years with the same accuracy as the $\mathrm{P}_{\mathrm{included}}$ model. Notably, there was an extremely large discrepancy when predicting the 2003-2004 flu season (4.4). Further testing to identify outlier signals showed that there may be consistently reduced performance. 10 week predictions provided the largest divergence in predictions.When two models were compared on the test data set containing weeks 540 - 806, a small signal was evident(4.3). However, the signal is very noisy.

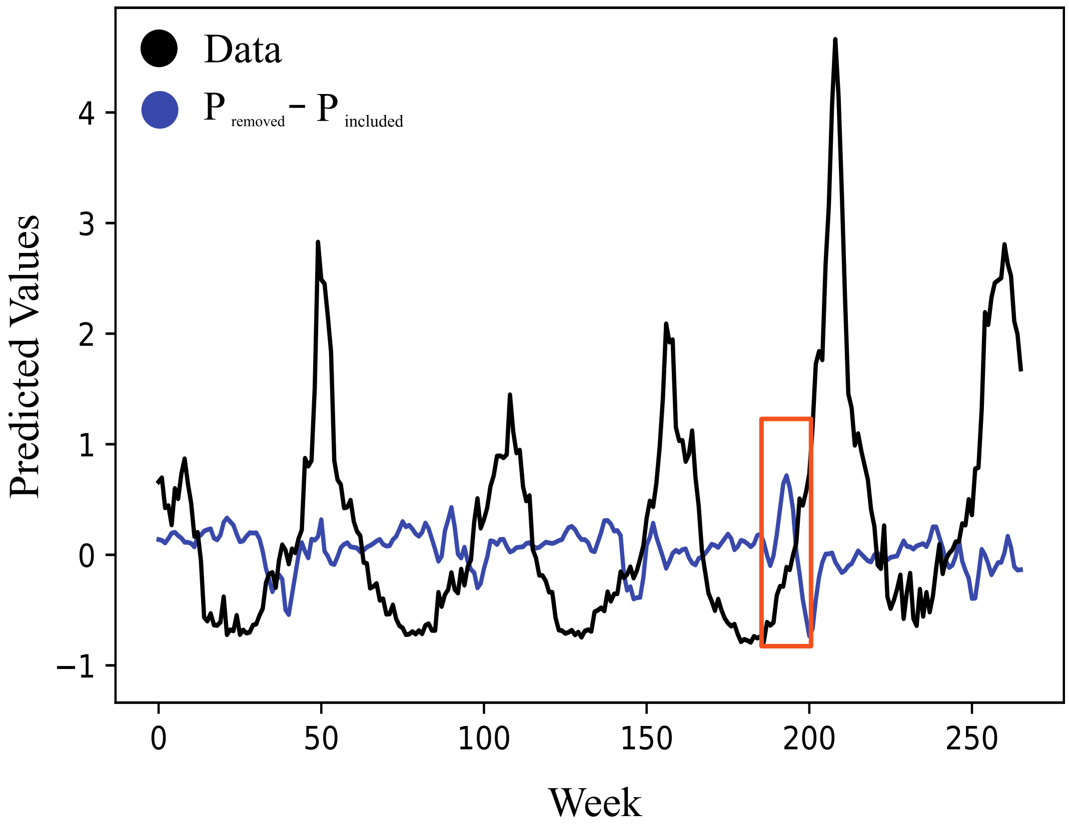

Figure 4.3 Two models trained on different data sets may provide a usable signal for identifying outlier flu years. Measuring divergence between two alternatively trained models provides a weak signal indicting possible outlier years. Week 10 predictions contained the largest error but also contained the largest systematic divergence, despite noise. Models trained on weeks 0 - 540 (approximate due to outlier removal) and tested on remaining weeks. Predicted values are weighted percent ILI. Data plotted has been standardized resulting in negative values on the y-axis.

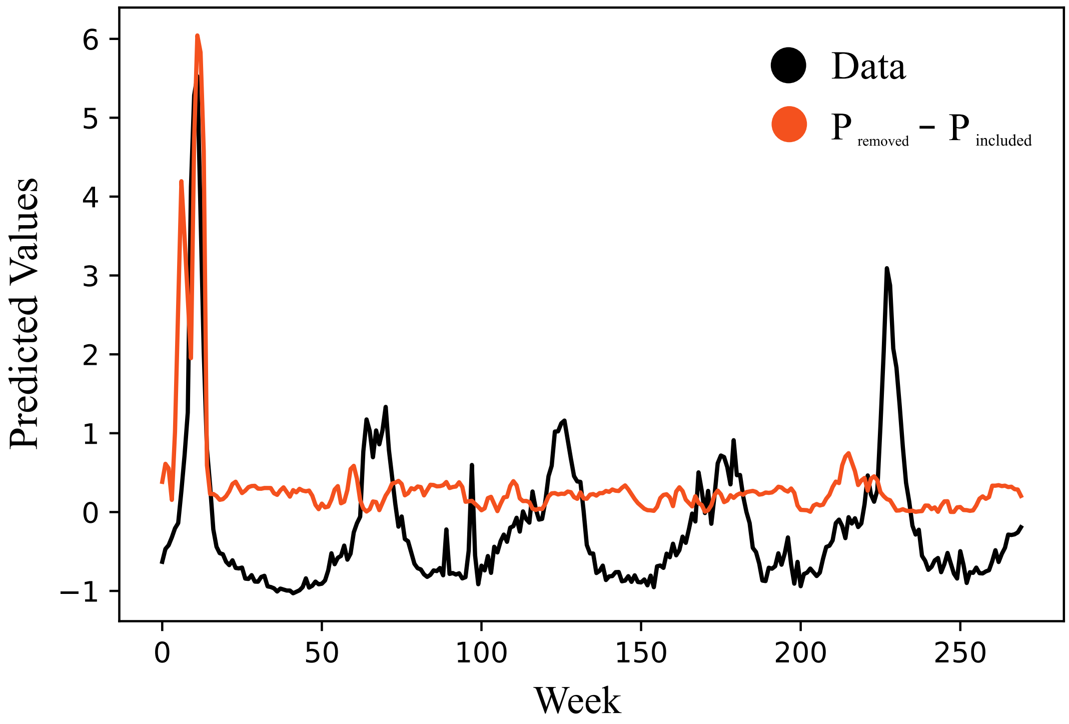

Figure 4.4 Removing outliers from training data set produces large prediction error in test outlier years. When trained on weeks 270 to 806 (approximately) models produce large prediction errors compared to models trained on complete data sets. Models trained on weeks 270 - 806 with and without outlier years. Predicted values are weighted percent ILI. Data plotted has been standardized resulting in negative values on the y-axis.

CHAPTER 5

DISCUSSION

Influenza produces seasonal outbreaks that have large economic and human costs. Currently, our best defense against seasonal outbreaks is widespread vaccination. However, despite advances in virology, epidemiology and immunology, a perfect influenza vaccine has eluded researchers. Additionally, major pandemic seasons can occur unexpectedly. As a result, predicting when and how any given flu season progresses is of the utmost importance. Early warning can allow development of targeted vaccines and health service preparation. To that end, a model predicting weekly percent ILI was developed.

The model made use of an LSTM-based neural network. By treating flu data available from the CDC as a time series, useful predictions were made. To increase predictive performance, climate and population data was added to the training set. By manipulating training sets and training variables, we were able to draw several conclusions.

5.1 Climate, Especially Temperature is an Important Predictor in Temperate Regions

Temperature was the strongest climate predictor that was used. Removing temperature data resulted in a sharp decline in predictive performance. This model suggests that there is at least a correlation between temperature and ILI rates. However, it provides no insight as to why this correlation exists. Precipitation was also a significant predictor, although to a lesser degree. This may be a result of precipitation's correlation with humidity. So, in this case, precipitation may be acting as a partial proxy for humidity data. Although the actual effect of relative humidity on influenza virus transmission has been contested, its usefulness as a predictor in modeling remains unchanged. Thus, the use of precipitation in place of less uniform humidity data may be of some benefit for future models. Notably, adding future climate data to the model greatly improves predictive performance. With unlimited future climate data, predictions may be pushed much further past the 10 week limit seen in this thesis. However,this is likely not realistic given current meteorological predictions.

Including climate data in influenza models provides a measurable increase in predictive power. However, climate data can be difficult to gather. In this case, select weather stations were used and the data averaged to create regional approximations. It is possible that if the model was applied to a smaller geographical area with more uniform weather, the predictive effect of climate data may be even greater. Future effort may be dedicated to collecting higher quality climate data from other sources. More granular data seemed to have a larger effect than less granular data. Weekly averages derived from daily data had a larger impact than weekly averages derived from monthly data. However, monthly data may be sufficient in some cases. Investigating the relationship between weekly influenza rates and the time-step of climate data may provide an insight into how and why climate effects influenza.

5.2 Population and Vaccination Data May Not Be Relevant to Modeling Using LSTM-Based Models

Removing population and vaccination data had no effect and increased performance, respectively. While this may indicate that these factors are not useful predictors, it is more likely that the data available was not sufficient. The data structure may have been inadequate to reveal underlying patterns. The population data used in this thesis was limited to regional total populations and the vaccination data was limited to national data. If more specific, granular data could be collected, it may be extremely useful in predicting influenza. However, due to the variability in flu vaccine effectiveness it is unlikely to be a useful predictive tool. Yearly effectiveness can only be calculated retrospectively. Despite the the lack of impact of population data, it would likely be much more valuable when adding a spacial dimension to the model, although that is outside of the scope of this thesis.

5.3 Training Sets Matter

Depending on the training set, the ability to predict changed. To an extent, more data provides better predictions. However, in a similarly tuned model, the optimal training set seemed to be about 10 years or about 540 weeks. This translates to about two-thirds of the total data set. Due to the time scale of yearly trends, some patterns may not be apparent in the limited data set. Whether or not the model is able to identify these hidden patterns is difficult to know. This is a potential downside of using a machine learning approach. For the most part, the model is very opaque. Without the ability to understand exactly what the model is learning, the best way to test generalizability is to rotate training data, taking different windows of data.

The inclusion or exclusion of certain years certainly affects the predictive power. This may indicate an underlying pattern that is better represented in some years rather than others. It may also indicate patterns that are better represented by sets of years, implying that the large sequence of years is as important as the granular data. Unfortunately, as mentioned previously, limited data prevents further investigation of this problem.

A specific example of variability in training data is the exclusion of the 2009 pandemic year. Models including this year in training data were able to better predict testing data than those without. This may indicate a robust pattern that emerges when the 2009 season is included. More data is needed to expand the scope of the model and understand more about the driving factors. Unfortunately, older CDC data is incomplete and more difficult to manage. Thus, previous pandemic years are not readily available for inclusion. A specific effort to collect and complete the data would be needed before it could be applied to this model. Of course data is continuously added as time advances. As subsequent years can be collected, patterns may emerge and the model can be reevaluated. However, observing the difference between models trained with and without outlier years may provide a useful tool for alternative approaches to prediction.

5.4 Outlier Years May Be Identified Using Prediction Divergence

Given the limits in predicting the myriad of variables that influence influenza trends, combined with the fact that many drivers of seasonal influenza are not well understood, accurate weekly predictions further than several weeks seems unlikely. Using the difference in predictive performance between models trained with and without outlier years may significantly extend the usefulness of this model. When compared to other forecasting techniques, such as those entered in the annual CDC Flu-Casting competition, this model performs well. However, no direct quantitative comparison was made in the course of this thesis.

The model put forth by this thesis is easily implemented and extendable to a variety of experiments. So training two parallel models is quick and requires only minor data modification. When comparing the results of alternatively trained models, systematic differences in predictive ability are observed. Specifically, years with unusually high percent ILI, when removed, cause poor prediction of high incidence years.

Limited examples of outlier years exist in the current data set. This limits further exploration of this phenomenon. Additionally, definition of outlier years was arbitrary. A peak standardized ILI of 4% was chosen. Years that fell slightly below this threshold may include useful data. With more data, trends may be more clear and categories may be easier to select. Another potential pitfall is the time-series nature of the data. Removing individual years may have unknown affects on training. However, given that the $\mathrm{P}_{\mathrm{removed}}$ model predicted all years except outlier years were predicted equally as well as the P_{included}, this is unlikely. Again, only a larger data set could answer this definitively. Despite these problems, the main advantage of this outlier detection approach is that is can function with the limited data available and make use of predictions that may be too inaccurate for direct use.

5.5 LSTM-Based Models Provide a Useful Tool for Influenza Prediction

Applying machine learning to biological and epidemiological questions is a relatively new approach. There has been limited examination of LSTM neural networks as the basis for influenza tracking models. Good one week predictions show that this approach is practical for so called now-casting. Also, using a variety of techniques, including recursive predictions, models can be stretched to predict to an indefinite point in the future. However, predictive performance plateaus between 5 and 10 weeks out and so is limited.

The primary advantage of this model is the straightforward architecture. It is small and does not require a vast amount of computational power, although it is much faster when GPU accelerated. Once the model has been designed and implemented, new data can be continuously fed. This model could be set up to automatically extract climate and influenza data in real time from various sources. A simple pre-processing pipeline would allow the data to be added seamlessly. This would allow relatively low effort predictions. Additionally, this architecture may be applied to a variety of locals. Further testing would be needed to confirm generalizability, though. If the model proves generalizable, it could provide a useful tool for modeling influenza for smaller organizations with limited resources. This model also provides a solid framework for future research. Training and prediction time is short, which allows rapid testing and on the fly modifications.

Overall, the effectiveness of LSTM-based models as a predictive tool is supported by the results of this thesis. While machine learning may act as a "black box" with opaque inner workings, continuing to apply it to biological questions can provide a useful practical tool as well as reveal previously unknown patterns in a system. Prediction of ILI trends using machine learning may have wider implications for epidemiology. This model and these techniques could be effectively applied to almost any infectious disease that acts in a time-dependant fashion. With minor modifications to data processing, the model may be applied to smaller time-scale outbreaks as long as previous data exists.

Diverging predictions can also provide a novel approach to compensating for limited data sets. The overall predictive accuracy does not need to be especially high. Instead, models must vary in predictive performance in a regular way. In cases where this holds true, this approach may provide a low-bar-of-entry approach for determining specific outbreak severity well in advance of other modeling techniques.

5.6 Future Research

Moving forward, this model can be further tuned. By manipulating the model, a better understanding of the underlying mechanisms may be had. The model may be simplified and studied in depth. Or the model may be used as a base for further expansion. For example, this model may be readily applied to an automatically updated web-based system. While this would not provide insight into the driving factors of influenza trends, it would provide a continuous stream of current data and updated models. Over the next several years, patterns may become more apparent.

Greater amounts of data are not always useful for increasing model accuracy, but better data could provide a measurable improvement. Climate data, especially, provides a large amount of room for improvement. Potentially using more stations, or better selected stations may provide better correlated data. Looking at areas with more uniform weather patterns could alleviate this issue. So, a future experiment may look at a single state, county or even city. Narrowing the area of interest may allow for better climate data.

Improving population data also provides room for further exploration. A study using the lessons learned from this model could add a spacial dimension to forecasting. By adding another dimension to the analysis, trends that were previously hidden may become apparent. Specifically, with the advent of effective image recognition, integration of heat-map type images could allow for straightforward addition of population density data as well as travel and immigration data. This type of modeling could be applied to almost any other transmissible disease.

In addition to better integrating population data to provide spacial understanding, flu sub-typing data may also be included. In this thesis, sub-type data was not included because data collection was sporadic and inconsistent year to year. More complete sub-typing data may lead to useful discoveries. Evaluation of this is outside the scope of this thesis. The model produced here does provide a usable framework for such an evaluation, though. Additionally, age data was not evaluated using this model, although the data was available and the architecture would allow for inclusion.

5.7 Final Thoughts

In conclusion, this thesis has shown that an LSTM-based model for predicting ILI trends is practical. The model produces usable results out to 10 weeks, further than required by the CDC competition on flu forecasting. Predictions past 10 weeks were not viable and so, by extending time of predictions, this model may reveal an outer limit for forecasting. Also, by expanding the data set, climate data has been confirmed as being a useful predictor. Specifically, this is the first LSTM-based model to incorporate climate, population and vaccination data. Furthermore, by manipulating the training data, systematic variations in predictions may be seen.

Using this, a simple system for identifying potential outlier flu years can be created. The work done for this thesis lays the foundation for a potentially novel approach to influenza forecasting.

APPENDIX A

SUPPLEMENTAL DATA, SAMPLE CODE & NETWORK ARCHITECTURES

This appendix contains select code, model architecture and links to complete data sets. The complete supplemental material can be found online.

A.1 Data

Complete data used is available online.

CDC influenza data: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html NOAA climate data: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets US Census population data: https://www.census.gov/data.html

A.2 Code Samples

This sample code contains the final model architecture used. The model was build using a Keras sequential model. The model is then wrapped in a simple function to compile and output the model for training. This model function is flexible enough to accept several data and label structures and so does not require modification to adjust time-step, output or other data structure changes. Code for previous iterations of this model is available online.

```

#define wrapper function for model creation
def build_modelE(data, labels):
        # data and labels must be pre-formatted as 3-D arrays with
    even
        # time-steps. Time steps corresponding to one week
        # one were chosen

        # take sample size for input. currently set to one
        # samples = data.shape[0]
        # time steps and features taken from data shape to
        # simplify building model
        # time_steps = data.shape[1]
        # features = data.shape[2]
        #output shape based on provided label data
        if len(labels.shape) == 2:
            label_shape = labels.shape[1]
        else:
            label_shape = 1

        # main model
        model = keras.Sequential([
                keras.layers.Bidirectional(layers.LSTM(500,
                return_sequences=True,
                batch_input_shape=((samples,
                time_steps,features))),
                merge_mode='concat'),
                keras.layers.Bidirectional(layers.LSTM(500,
                dropout=.3,
                return_sequences=True),
                merge_mode='concat'),
                keras.layers.Bidirectional(layers.LSTM(500,
                dropout=.3),
                merge_mode='concat'),
                keras.layers.Dense(label_shape)
        ])
        #uses mean absolute error as loss function
        model.compile(loss='mse',
                      optimizer=tf.keras.optimizers.Adam(),
                      metrics=['mae','mse', R_square])
        return model
```

Listing A.1 Final model architecture.

The code section below contains the function that allows for 10 week predictions with a 4 week time lag. This was one of the main functions used to evaluate the models. Complete code, including wrapper functions for other prediction strategies, is available online.

```

#
def predict_future_t4(model, data):
    # recursive prediction by week with 4 week lag

    # model - trained model to be used to make predictions.
  Should
    # predict 1 week forward and contain 4 week time lag

    # data - testing data. This data should be pre-formatted in
   the same
    # way as the training data



    # initialize numpy array to hold predictions
    # full_predictions = np.empty((data.shape[0],10))

    # main loop of prediction
    # loops predictions over each time step (in this case each week)
    for j in range(0,data.shape[0]):


        # initialize a numpy array of zeros to
        # store predictions
        predictions = np.zeros((10,1))

        # copy data to edit
        edit_data = data.copy()

        # secondary loop that predicts 10 weeks out from each time
        # step
        for i in range(j,j+10):
            #start with a 2 week slice of data
            #grow as more is predicted
            base_week = edit_data[:i+1]

            #generate prediction on data
            predict_week = model.predict(base_week)

            #loop predictions back into data
            #[row, 3rd dimension == 0 , column]

            #t-1
            if i+1<data.shape[0]:
                edit_data[i+1,0,41]=predict_week[-1]
            #t-2
            if i+2<data.shape[0] & i-j>=1:
                edit_data[i+2,0,28]=predict_week[-1]
            #t-3
            if i+3<data.shape[0] & i-j>=2:
                edit_data[i+3,0,15]=predict_week[-1]
            #t-4
            if i+4<data.shape[0] & i-j>=3:
                edit_data[i+4,0,2]=predict_week[-1]

            #add prediction to prediction list
            predictions[i-j,] = predict_week[-1,]

        predictions = predictions.flatten()
        predictions = np.reshape(predictions, (1,10))

        # simple add-on to print status of predictions
        j_number = j+1
        print("{j_num}/{number}".format(j_num=j_number,
        number=data.shape[0]),
        end= " ", flush=True)
        full_predictions[j,] = predictions[0,]

    return full_predictions
```

Listing A.2 Select code for recursive and standard predictive functions.

A.3 Model Architecture

```
def build_modelE(data, labels):
      #take sample size for input. currently set to one
      samples = data.shape[0]
      # time steps and features taken from data shape to simplify
      # building model
      time_steps = data.shape[1]
      features = data.shape[2]
      #output shape
      if len(labels.shape) == 2:
            label_shape = labels.shape[1]
      else:
            label_shape = 1


      model = keras.Sequential([
                keras.layers.Bidirectional(layers.LSTM(500,
                return_sequences=True,
                    batch_input_shape=((samples,time_steps,
   features))),
                    merge_mode='concat'),
                keras.layers.Bidirectional(layers.LSTM(500,
                    dropout=.25, return_sequences=True),
                    merge_mode='concat'),
                keras.layers.Bidirectional(layers.LSTM(500,
                    dropout=.25),merge_mode='concat'),
                keras.layers.Dense(label_shape)
                ])

```

```
    #uses mean absolute error as loss function
    model.compile(loss='mse',
                  optimizer=tf.keras.optimizers.Adam(),
                  metrics=['mae','mse', R_square])
    return model
```

Listing A.3 Code for final model architecture.

BIBLIOGRAPHY

[1] Centers for disease control and prevention.

[2] *Assessment of the Effectiveness of the 2003–04 Influenza Vaccine Among Children and Adults — Colorado, 2003.* Center for Disease Control, 2004.

[3] Types of influenza viruses. *Center for Disease Control*, Sep 2017. https://www.cdc.gov/flu/about/viruses/types.htm.

[4] Estimated influenza illnesses, medical visits, hospitalizations, and deaths in the united states 2017–2018 influenza season. *Center for Disease Control*, Dec 2018. https://www.cdc.gov/flu/about/burden/2017-2018.htm.

[5] Influenza (seasonal). https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal), Nov 2018.

[6] Key facts about influenza (flu). *Center for Disease Control*, Apr 2018. https://www.cdc.gov/flu/about/keyfacts.htm.

[7] Gillian M Air and W Graeme Laver. The neuraminidase of influenza virus. *Proteins: Structure, Function, and Bioinformatics*, 6(4):341–356, 1989.

[8] Jacob Bock Axelsen, Rami Yaari, Bryan T Grenfell, and Lewi Stone. Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. *Proceedings of the National Academy of Sciences*, 111(26):9538–9542, 2014.

[9] Edward A Belongia, Burney A Kieke, James G Donahue, Laura A Coleman, Stephanie A Irving, Jennifer K Meece, Mary Vandermause, Stephen Lindstrom, Paul Gargiullo, and David K Shay. Influenza vaccine effectiveness in wisconsin during the 2007–08 season: comparison of interim and final results. *Vaccine*, 29(38):6558–6563, 2011.

[10] Edward A Belongia, Burney A Kieke, James G Donahue, Robert T Greenlee, Amanda Balish, Angie Foust, Stephen Lindstrom, and David K Shay. Effectiveness of inactivated influenza vaccines varied substantially with antigenic match from the 2004–2005 season to the 2006–2007 season. *The Journal of infectious diseases*, 199(2):159–167, 2009.

[11] Nicole M Bouvier. Animal models for influenza virus transmission studies: a historical perspective. *Current opinion in virology*, 13:101–108, 2015.

[12] Thomas J Braciale, Jie Sun, and Taeg S Kim. Regulating the adaptive immune response to respiratory virus infection. *Nature Reviews Immunology*, 12(4):295, 2012.

[13] Gabrielle Brankston, Leah Gitterman, Zahir Hirji, Camille Lemieux, and Michael Gardam. Transmission of influenza a in human beings. *The Lancet infectious diseases*, 7(4):257–265, 2007.

[14] Abraham L Brass, I-Chueh Huang, Yair Benita, Sinu P John, Manoj N Krishnan, Eric M Feeley, Bethany J Ryan, Jessica L Weyer, Louise Van Der Weyden, Erol Fikrig, et al. The ifitm proteins mediate cellular resistance to influenza a h1n1 virus, west nile virus, and dengue virus. *Cell*, 139(7):1243–1254, 2009.

[15] JJ Cannell, R Vieth, JC Umhau, MF Holick, WB Grant, S Madronich, CF Garland, and E Giovannucci. Epidemic influenza and vitamin d. *Epidemiology & Infection*, 134(6):1129–1140, 2006.

[16] Gongbo Chen, Wenyi Zhang, Shanshan Li, Yongming Zhang, Gail Williams, Rachel Huxley, Hongyan Ren, Wei Cao, and Yuming Guo. The impact of ambient fine particles on influenza transmission and the modification effects of temperature in china: a multi-city study. *Environment international*, 98:82–88, 2017.

[17] Rubing Chen and Edward C Holmes. Avian influenza virus exhibits rapid evolutionary dynamics. *Molecular biology and evolution*, 23(12):2336–2341, 2006.

[18] Richard A Cone. Barrier properties of mucus. *Advanced drug delivery reviews*, 61(2):75–85, 2009.

[19] Gina M Conenello, Dmitriy Zamarin, Lucy A Perrone, Terrence Tumpey, and Peter Palese. A single mutation in the pb1-f2 of h5n1 (hk/97) and 1918 influenza a viruses contributes to increased virulence. *PLoS pathogens*, 3(10):e141, 2007.

[20] Benjamin D Dalziel, Stephen Kissler, Julia R Gog, Cecile Viboud, Ottar N Bjørnstad, C Jessica E Metcalf, and Bryan T Grenfell. Urbanization and humidity shape the intensity of influenza epidemics in us cities. *Science*, 362(6410):75–79, 2018.

[21] Ethan R Deyle, M Cyrus Maher, Ryan D Hernandez, Sanjay Basu, and George Sugihara. Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences*, 113(46):13081–13086, 2016.

[22] Andrea Di Pietro, Anna Kajaste-Rudnitski, Alexandra Oteiza, Lucia Nicora, Greg J Towers, Nadir Mechti, and Elisa Vicenzi. Trim22 inhibits influenza a virus infection by targeting the viral nucleoprotein for degradation. *Journal of virology*, 87(8):4523–4533, 2013.

[23] Jonathan Dushoff, Joshua B Plotkin, Simon A Levin, and David JD Earn. Dynamical resonance can account for seasonality of influenza epidemics. *Proceedings of the National Academy of Sciences*, 101(48):16915–16916, 2004.

[24] Epidemiology, Research Statistics Unit, and Health Education Division. Trends in pneumonia and influenza morbidity and mortality, 2015.

[25] Ryan S Ference, James A Leonard, and Howard D Stupak. Physiologic model for seasonal patterns in flu transmission. *The Laryngoscope*, 2019.

[26] Brendan Flannery, Jessie R Chung, Swathi N Thaker, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Manjusha Gaglani, Kempapura Murthy, Richard K Zimmerman, et al. Interim estimates of 2016–17 seasonal influenza vaccine effectivenessunited states, february 2017. *MMWR. Morbidity and mortality weekly report*, 66(6):167, 2017.

[27] Manjusha Gaglani, Jessica Pruszynski, Kempapura Murthy, Lydia Clipper, Anne Robertson, Michael Reis, Jessie R Chung, Pedro A Piedra, Vasanthi Avadhanula, Mary Patricia Nowalk, et al. Influenza vaccine effectiveness against 2009 pandemic influenza a (h1n1) virus differed by vaccine type during 2013–2014 in the united states. *The Journal of infectious diseases*, 213(10):1546–1556, 2016.

[28] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[29] Marc P Girard, John S Tam, Olga M Assossou, and Marie Paule Kieny. The 2009 a (h1n1) influenza virus pandemic: A review. *Vaccine*, 28(31):4895–4902, 2010.

[30] Kyra H Grantz, Madhura S Rane, Henrik Salje, Gregory E Glass, Stephen E Schachterle, and Derek AT Cummings. Disparities in influenza mortality and transmission related to sociodemographic factors within chicago in the pandemic of 1918. *Proceedings of the National Academy of Sciences*, 113(48):13839–13844, 2016.

[31] Marie R Griffin, Arnold S Monto, Edward A Belongia, John J Treanor, Qingxia Chen, Jufu Chen, H Keipp Talbot, Suzanne E Ohmit, Laura A Coleman, Gerry Lofthus, et al. Effectiveness of non-adjuvanted pandemic influenza a vaccines for preventing pandemic influenza acute respiratory illness visits in 4 us communities. *PloS one*, 6(8):e23085, 2011.

[32] Zhenhong Guan, Di Liu, Shuofu Mi, Jie Zhang, Qinong Ye, Ming Wang, George F Gao, and Jinghua Yan. Interaction of hsp40 with influenza virus m2 protein: implications for pkr signaling pathway. *Protein & cell*, 1(10):944–955, 2010.

[33] Y Guo, M Wang, FG Jin, P Wang, and JM Zhu. Influenza ecology in china. *The origin of pandemic influenza viruses. Elsevier Science Publishing, Inc., New York*, pages 211–220, 1983.

[34] Frederick G Hayden and Menno D de Jong. Human influenza: Pathogenesis, clinical features, and management. In RD Webtser, AS Monto, TJ Braciale, and RA Lamb, editors, *Textbook of Influenza*, chapter 24, pages 373–391. John Wiley & Sons, Ltd West Sussex:, 2013.

[35] Cheng-Qiang He, Zhi-Xun Xie, Guan-Zhu Han, Jian-Bao Dong, Dong Wang, Jia-Bo Liu, Le-Yuan Ma, Xiao-Fei Tang, Xi-Ping Liu, Yao-Shan Pang, et al. Homologous recombination as an evolutionary force in the avian influenza a virus. *Molecular biology and evolution*, 26(1):177–187, 2008.

[36] Hans Peter Hefti, Michael Frese, Heinrich Landis, Claudio Di Paolo, Adriano Aguzzi, Otto Haller, and Jovan Pavlovic. Human mxa protein protects mice lacking a functional alpha/beta interferon system against la crosse virus and other lethal viral infections. *Journal of virology*, 73(8):6984–6991, 1999.

[37] Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997.

[38] Aleksandr Ianevski, Eva Zusinaite, Nastassia Shtaida, Hannimari Kallio-Kokko, Miia Valkonen, Anu Kantele, Kaidi Telling, Irja Lutsar, Pille Letjuka, Natalja Metelitsa, et al. Low temperature and low uv indexes correlated with peaks of influenza virus activity in northern europe during 2010–2018. *Viruses*, 11(3):207, 2019.

[39] Akiko Iwasaki and Padmini S Pillai. Innate immunity to influenza virus infection. *Nature Reviews Immunology*, 14(5):315, 2014.

[40] Charlotte Jackson, Emilia Vynnycky, and Punam Mangtani. The relationship between school holidays and transmission of influenza in england and wales. *American journal of epidemiology*, 184(9):644–651, 2016.

[41] Michael L Jackson, Jessie R Chung, Lisa A Jackson, C Hallie Phillips, Joyce Benoit, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Manjusha Gaglani, et al. Influenza vaccine effectiveness in the united states during the 2015–2016 season. *New England Journal of Medicine*, 377(6):534–543, 2017.

[42] Edwin D Kilbourne. Influenza pandemics of the 20th century. *Emerging infectious diseases*, 12(1):9, 2006.

[43] Ben Killingley and Jonathan Nguyen-Van-Tam. Routes of influenza transmission. *Influenza and other respiratory viruses*, 7:42–51, 2013.

[44] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[45] JHCM Kreijtz, RAM Fouchier, and GF Rimmelzwaan. Immune responses to influenza virus infection. *Virus research*, 162(1-2):19–30, 2011.

[46] RM Krug and E Fodor. The virus genome and its replication. In RD Webtser, AS Monto, TJ Braciale, and RA Lamb, editors, *Textbook of Influenza*, chapter 3, pages 57–66. John Wiley & Sons, Ltd West Sussex:, 2013.

[47] Jasmin S Kutter, Monique I Spronken, Pieter L Fraaij, Ron AM Fouchier, and Sander Herfst. Transmission routes of respiratory viruses among humans. *Current opinion in virology*, 28:142–151, 2018.

[48] Soon-Bark Kwon, Jaehyung Park, Jaeyoun Jang, Youngmin Cho, Duck-Shin Park, Changsoo Kim, Gwi-Nam Bae, and Am Jang. Study on the initial velocity distribution of exhaled air from coughing and speaking. *Chemosphere*, 87(11):1260–1264, 2012.

[49] Robert A Lamb and Purnell W Choppin. Segment 8 of the influenza virus genome is unique in coding for two polypeptides. *Proceedings of the National Academy of Sciences*, 76(10):4908–4912, 1979.

[50] Robert A Lamb, Leslie J Holsinger, and Lawrence H Pinto. 16 the influenza a virus m2 ion channel protein and its role in the influenza virus life cycle. *Cold Spring Harbor Monograph Archive*, 28:303–321, 1994.

[51] Robert A Lamb, Suzanne L Zebedee, and Christopher D Richardson. Influenza virus m2 protein is an integral membrane protein expressed on the infected-cell surface. *Cell*, 40(3):627–633, 1985.

[52] William G Lindsley, John D Noti, Francoise M Blachere, Robert E Thewlis, Stephen B Martin, Sreekumar Othumpangat, Bahar Noorbakhsh, William T Goldsmith, Abhishek Vishnu, Jan E Palmer, et al. Viable influenza a virus in airborne particles from human coughs. *Journal of occupational and environmental hygiene*, 12(2):107–113, 2015.

[53] Liyuan Liu, Meng Han, Yiyun Zhou, and Yan Wang. Lstm recurrent neural networks for influenza trends prediction. In *International Symposium on Bioinformatics Research and Applications*, pages 259–264. Springer, 2018.

[54] Eric Lofgren, Nina H Fefferman, Yuri N Naumov, Jack Gorski, and Elena N Naumova. Influenza seasonality: underlying causes and modeling theories. *Journal of virology*, 81(11):5429–5436, 2007.

[55] Anice C Lowen, Samira Mubareka, John Steel, and Peter Palese. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151, 2007.

[56] Anice C Lowen and John Steel. Roles of humidity and temperature in shaping influenza seasonality. *Journal of virology*, 88(14):7692–7695, 2014.

[57] Paul MV Martin and Estelle Martin-Granel. 2,500-year evolution of the term epidemic. *Emerging infectious diseases*, 12(6):976, 2006.

[58] Mikhail N Matrosovich, Tatyana Y Matrosovich, Thomas Gray, Noel A Roberts, and Hans-Dieter Klenk. Neuraminidase is important for the initiation of influenza virus infection in human airway epithelium. *Journal of virology*, 78(22):12665–12667, 2004.

[59] Julie McAuley, Brad Gilbertson, Sanja Trifkovic, Lorena Elizabeth Brown, and Jennifer McKimm-Breschkin. Influenza virus neuraminidase structure and functions. *Frontiers in microbiology*, 10:39, 2019.

[60] Huong Q McLean, Mark G Thompson, Maria E Sundaram, Burney A Kieke, Manjusha Gaglani, Kempapura Murthy, Pedro A Piedra, Richard K Zimmerman, Mary Patricia Nowalk, Jonathan M Raviotta, et al. Influenza vaccine effectiveness in the united states during 2012–2013: variable protection by age and virus type. *The Journal of infectious diseases*, 211(10):1529–1540, 2014.

[61] Arnold S Monto, Stefan Gravenstein, Michael Elliott, Michael Colopy, and Jo Schweinle. Clinical signs and symptoms predicting influenza infection. *Archives of internal medicine*, 160(21):3243–3247, 2000.

[62] Arnold S Monto and Robert G Webster. Influenza pandemics: History and lessons learned. In RD Webtser, AS Monto, TJ Braciale, and RA Lamb, editors, *Textbook of Influenza*, chapter 2, pages 20–33. John Wiley & Sons, Ltd West Sussex:, 2013.

[63] Debiprosad Naya, Sakar Shivakoti, Rilwan A. Balogun, Gwendolyn Lee1, and Z. Hong Zhou. Structure, disassembly, assembly, and budding of influenza viruses. In RD Webtser, AS Monto, TJ Braciale, and RA Lamb, editors, *Textbook of Influenza*, chapter 3, pages 37–56. John Wiley & Sons, Ltd West Sussex:, 2013.

[64] Debi P Nayak, Rilwan A Balogun, Hiroshi Yamada, Z Hong Zhou, and Subrata Barman. Influenza virus morphogenesis and budding. *Virus research*, 143(2):147–161, 2009.

[65] Debi P Nayak and Eric K-W Hui. The role of lipid microdomains in virus biology. In *Membrane Dynamics and Domains*, pages 443–491. Springer, 2004.

[66] Gabriele Neumann and Yoshihiro Kawaoka. Transmission of influenza a viruses. *Virology*, 479:234–246, 2015.

[67] Sophia Ng, Roger Lopez, Guillermina Kuan, Lionel Gresh, Angel Balmaseda, Eva Harris, and Aubree Gordon. The timeline of influenza virus shedding in children and adults in a household transmission study of influenza in managua, nicaragua. *The Pediatric infectious disease journal*, 35(5):583, 2016.

[68] Suzanne E Ohmit, Mark G Thompson, Joshua G Petrie, Swathi N Thaker, Michael L Jackson, Edward A Belongia, Richard K Zimmerman, Manjusha Gaglani, Lois Lamerato, Sarah M Spencer, et al. Influenza vaccine effectiveness in the 2011–2012 season: protection against each circulating virus and the effect of prior vaccination on estimates. *Clinical infectious diseases*, 58(3):319–327, 2013.

[69] Robert E O'Neill, Julie Talon, and Peter Palese. The influenza virus nep (ns2 protein) mediates the nuclear export of viral ribonucleoproteins. *The EMBO journal*, 17(1):288–296, 1998.

[70] Jeffrey D Parvin, A Moscona, WT Pan, JM Leider, and P Palese. Measurement of the mutation rates of animal viruses: influenza a virus and poliovirus type 1. *Journal of virology*, 59(2):377–383, 1986.

[71] Catharine I Paules and Anthony S Fauci. Influenza vaccines: good, but we can do better. *The Journal of infectious diseases*, 219(Supplement_1):S1–S4, 2019.

[72] Joseph Sriyal Malik Peiris, Chung Yan Cheung, Connie Yin Hung Leung, and John Malcolm Nicholls. Innate immune responses to influenza a h5n1: friend or foe? *Trends in immunology*, 30(12):574–584, 2009.

[73] JS Malik Peiris, Leo LM Poon, and Yi Guan. Emergence of a novel swine-origin influenza a virus (s-oiv) h1n1 virus in humans. *Journal of Clinical Virology*, 45(3):169–173, 2009.

[74] Nicholas G Reich, Logan Brooks, Spencer Fox, Sasikiran Kandula, Craig McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa Yamana, et al. Forecasting seasonal influenza in the us: A collaborative multi-year, multi-model assessment of forecast performance. *bioRxiv*, page 397190, 2018.

[75] Patricia Resa-Infante, Núria Jorba, Rocio Coloma, and Juan Ortín. The influenza virus rna synthesis machine: advances in its structure and function. *RNA biology*, 8(2):207–215, 2011.

[76] Melissa A Rolfes, Brendan Flannery, Jessie R Chung, Alissa OHalloran, Shikha Garg, Edward A Belongia, Manjusha Gaglani, Richard K Zimmerman, Michael L Jackson, Arnold S Monto, et al. Effects of influenza vaccination in the united states during the 2017–2018 influenza season. *Clinical Infectious Diseases*, 2019.

[77] Jeremy S Rossman and Robert A Lamb. Influenza virus assembly and budding. *Virology*, 411(2):229–236, 2011.

[78] Tasleem Samji. Influenza a: understanding the viral life cycle. *The Yale journal of biology and medicine*, 82(4):153, 2009.

[79] Jason R Schnell and James J Chou. Structure and mechanism of the m2 proton channel of influenza a virus. *Nature*, 451(7178):591, 2008.

[80] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[81] Jeffrey Shaman, Virginia E Pitzer, Cécile Viboud, Bryan T Grenfell, and Marc Lipsitch. Absolute humidity and the seasonal onset of influenza in the continental united states. *PLoS biology*, 8(2):e1000316, 2010.

[82] Wenhan Shao, Xinxin Li, Mohsan Goraya, Song Wang, and Ji-Long Chen. Evolution of influenza a virus by mutation and re-assortment. *International journal of molecular sciences*, 18(8):1650, 2017.

[83] GI Shapiro, T Gurney, and RM Krug. Influenza virus gene expression: control mechanisms at early and late times of infection and nuclear-cytoplasmic transport of virus-specific rnas. *Journal of virology*, 61(3):764–773, 1987.

[84] Richard E Shope. Swine influenza: Iii. filtration experiments and etiology. *Journal of Experimental Medicine*, 54(3):373–385, 1931.

[85] RICHARD E Shope. Swine influenza. *Disease of swine. Iowa State University Press, Ames*, pages 81–91, 1958.

[86] John J Skehel and Don C Wiley. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual review of biochemistry*, 69(1):531–569, 2000.

[87] Radina P Soebiyanto, Farida Adimi, and Richard K Kiang. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PloS one*, 5(3):e9450, 2010.

[88] RJ Sugrue and AJ Hay. Structural characteristics of the m2 protein of influenza a viruses: Evidence that it forms a tetrameric channe. *Virology*, 180(2):617–624, 1991.

[89] James Tamerius, Martha I Nelson, Steven Z Zhou, Cécile Viboud, Mark A Miller, and Wladimir J Alonso. Global influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental health perspectives*, 119(4):439–445, 2010.

[90] James D Tamerius, Jeffrey Shaman, Wladmir J Alonso, Kimberly Bloom-Feshbach, Christopher K Uejio, Andrew Comrie, and Cécile Viboud. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS pathogens*, 9(3):e1003194, 2013.

[91] Shin-ichi Tamura and Takeshi Kurata. Defense mechanisms against influenza virus infection in the respiratory tract mucosa. *Jpn J Infect Dis*, 57(6):236–47, 2004.

[92] Raymond Tellier. Review of aerosol transmission of influenza a virus. *Emerging infectious diseases*, 12(11):1657, 2006.

[93] John J Treanor, H Keipp Talbot, Suzanne E Ohmit, Laura A Coleman, Mark G Thompson, Po-Yung Cheng, Joshua G Petrie, Geraldine Lofthus, Jennifer K Meece, John V Williams, et al. Effectiveness of seasonal influenza vaccines in the united states during a season with circulation of all three vaccine strains. *Clinical infectious diseases*, 55(7):951–959, 2012.

[94] JN Varghese, WG Laver, and Peter M Colman. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 å resolution. *Nature*, 303(5912):35, 1983.

[95] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one*, 12(12):e0188941, 2017.

[96] Robert G Webster, William J Bean, Owen T Gorman, Thomas M Chambers, and Yoshihiro Kawaoka. Evolution and ecology of influenza a viruses. *Microbiology and molecular biology reviews*, 56(1):152–179, 1992.

[97] Robert G Webster, Maya Yakhno, Virginia S Hinshaw, William J Bean, and K Copal Murti. Intestinal influenza: replication and characterization of influenza viruses in ducks. *Virology*, 84(2):268–278, 1978.

[98] W Weis, JH Brown, S Cusack, JC Paulson, JJ Skehel, and DC Wiley. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature*, 333(6172):426, 1988.

[99] Don C Wiley and John J Skehel. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual review of biochemistry*, 56(1):365–394, 1987.

[100] Ian A Wilson, John J Skehel, and DC Wiley. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 å resolution. *Nature*, 289(5796):366, 1981.

[101] Helen M. Wise, Agnes Foeglein, Jiechao Sun, Rosa Maria Dalton, Sheetal Patel, Wendy Howard, Emma C. Anderson, Wendy S. Barclay, and Paul Digard. A complicated message: Identification of a novel pb1-related protein translated from influenza a virus segment 2 mrna. *Journal of Virology*, 83(16):8021–8031, 2009.

[102] Qiaozhen Ye, Robert M Krug, and Yizhi Jane Tao. The mechanism by which influenza a virus nucleoprotein forms oligomers and binds rna. *Nature*, 444(7122):1078, 2006.

[103] Maria C Zambon. The pathogenesis of influenza in humans. *Reviews in medical virology*, 11(4):227–241, 2001.

[104] Suzanne L Zebedee and Robert A Lamb. Influenza a virus m2 protein: monoclonal antibody restriction of virus growth and detection of m2 in virions. *Journal of virology*, 62(8):2762–2772, 1988.

[105] Jie Zhang and Kazumitsu Nawata. A comparative study on predicting influenza outbreaks. *Bioscience trends*, 2017.

[106] Richard K Zimmerman, Mary Patricia Nowalk, Jessie Chung, Michael L Jackson, Lisa A Jackson, Joshua G Petrie, Arnold S Monto, Huong Q McLean, Edward A Belongia, Manjusha Gaglani, et al. 2014–2015 influenza vaccine effectiveness in the united states by vaccine type. *Clinical Infectious Diseases*, page 635, 2016.