SANDIA REPORT
SAND20XX-XXXX
Printed December 2018

Sandia National Laboratories

Biologically inspired approaches for biosurveillance anomaly detection and data fusion

Patrick Finley, Drew Levin, Tatiana Flanagan, Walt Beyeler, Michael Mitchell, Jaideep Ray, Melanie Moses, Stephanie Forrest

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

ABSTRACT

This study developed and tested biologically inspired computational methods to detect anomalous signals in data streams that could indicate a pending outbreak or bio-weapon attack. Current large-scale biosurveillance systems are plagued by two principal deficiencies: (1) timely detection of disease-indicating signals in noisy data and (2) anomaly detection across multiple channels. Anomaly detectors and data fusion components modeled after human immune system processes were tested against a variety of natural and synthetic surveillance datasets. A pilot scale immune-system-based biosurveillance system performed at least as well as traditional statistical anomaly detection data fusion approaches. Machine learning approaches leveraging Deep Learning recurrent neural networks were developed and applied to challenging unstructured and multimodal health surveillance data. Within the limits imposed of data availability, both immune systems and deep learning methods were found to improve anomaly detection and data fusion performance for particularly challenging data subsets.

ACKNOWLEDGEMENTS

The authors acknowledge the close collaboration of Scott Lee, Jason Thomas, and Chad Heilig from the US Centers for Disease Control (CDC) in this effort. De-identified biosurveillance data provided by Ken Jeter of the New Mexico Department of Health proved to be an important contribution to our work. Discussions with members of the International Society of Disease Surveillance helped the researchers focus on questions relevant to practicing public health professionals. Funding for this work was provided by Sandia National Laboratories' Laboratory Directed Research and Development program.

CONTENTS

1. Introduction 10
2. Biosurveillance data 11
3. Immune System anomaly detection and data fusion 12
   - 3.1. Dendritic Cell Algorithm 12
     - 3.1.1. Algorithm Description 12
     - 3.1.2. Experimental 13
   - 3.2. Negative Selection 13
     - 3.2.1. Algorithm Description 14
     - 3.2.2. Experimental 14
   - 3.3. Negative Selection Data Fusion 15
     - 3.3.1. Experimental 16
   - 3.4. Biosurveillance Network Topology 17
4. Neural Network anomaly detection and data fusion 19
   - 4.1. Traditional Methods 19
     - 4.1.1. Anomaly Detection 19
     - 4.1.2. Data Fusion 20
   - 4.2. Biosurveillance Challenges 20
     - 4.2.1. Multiple data types 20
     - 4.2.2. Temporal lags 21
     - 4.2.3. Missing data 21
     - 4.2.4. Pattern Recognition 21
     - 4.2.5. Fusion Order 21
     - 4.2.6. Deep Learning Alternatives 22
     - 4.2.7. Sequence Analysis 22
     - 4.2.8. Time-series Analysis 23
     - 4.2.9. Anomaly Detection 23
     - 4.2.10. Data fusion 24
   - 4.3. Applying Deep Learning to Biosurveillance 24
     - 4.3.1. Missing and Aperiodic Data 24
     - 4.3.2. Early vs Late Fusion 24
     - 4.3.3. Correlated Temporal Lags 24
     - 4.3.4. Pattern Recognition 25
     - 4.3.5. Summary 25
   - 4.4. Experimental Evaluation of Deep Learning Approaches 25
     - 4.4.1. Data Fusion Approach 25
   - 4.5. Experimental 26
     - 4.5.1. Resolving Temporal Offsets 26
     - 4.5.2. Mixed-Mode Datasets 27
       - 4.5.2.1. Biosurveillance Data Set 28
       - 4.5.2.2. Single Mode Text Classification 28
       - 4.5.2.3. Multimodal Data Fusion 28
   - 4.6. Future Research Directions 30
     - 4.6.1. Early vs Late Fusion 30
     - 4.6.2. Tuning for Specific Data Types 30
5. Discussion 32

5.1. Immune system analogs 32
6. Conclusions 34
6.1. Immune system analogs 34
6.2. Deep Neural Networks 34
6.3. Future Directions 35

LIST OF FIGURES

Figure 1. Naive DCA implementation. 13
Figure 2. Negative selection anomaly detector sensitivity and specificity. S 15
Figure 3. Application of negative selection for anomaly detection on a sample dataset. 16
Figure 4. Conceptual models of centralized and distributed national biosurveillance networks. 17
Figure 5. MDL data fusion results. 27
Figure 6. Schematic diagram of MDL data fusion 27
Figure 7. Chief complaint classification performance measured by ROC AUC and PRC. 29
Figure 8. Schematic illustration of early, late and hybrid data fusion. 30

LIST OF TABLES

Table 1. Accuracy of predicted discharge codes derived from text classification 29

This page left blank

EXECUTIVE SUMMARY

Biosurveillance systems collect and analyze vast quantities of diverse real-time data from to provide the Nation with advance warning of disease outbreak or bioweapons attack. The daily volume of monitored information from hospital admissions, emergency responders, drug purchases, and social-media search patterns is huge and growing rapidly as new sources come on line. Detection of bioweapon and disease-outbreak signals in these large noisy data streams challenges traditional statistics-based algorithms in current use. Scientists from Sandia National Laboratories (SNL), University of New Mexico (UNM), and US Centers for Disease Control (CDC) have applied advanced data analytics concepts to address these pressing national scale biosurveillance data-fusion and anomaly-recognition gaps. Researchers have developed a broad array of classifiers including artificial immune system models, traditional machine learning methods, and deep learning neural networks and tested their performance against production biosurveillance data sets and synthetic data sets. These methods show tremendous potential for detecting possibly devastating outbreaks earlier and with greater reliability. While some of the new methods remain more data-hungry and compute-intensive than traditional statistical approaches, additional algorithm tuning and new high performance computing architectures may enable the new methods to improve biosecurity readiness over the long term.

This report summarizes many statistical and computer science studies designed to explore novel methods to improve large-scale biosurveillance system performance. Technical details of the work summarized in this report are contained in individual publications and technical reports cited in the reference section.

ACRONYMS AND DEFINITIONS

Abbreviation	Definition
AD	Anomaly Detection
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
BOW	Bag of Words
C2W	Character to Word
CB	Chemical/Biological
CNN	Convolutional Neural Network
CUSUM	Cumulative Sum
DL	Deep Learning
DOD	Department of Defense
DTRA	Defense Threat Reduction Agency
DUA	Data Use Agreement
ED	Emergency Department
EHR	Electronic Health Record
EWMA	Exponential Weighted Moving Average
ICD	International Classification of Disease
IDF	Inverse Document Frequency
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MDL	Multimodal Deep Learning
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NKP	Natural Language Processing
NN	Neural Network
PII	Personally Identifiable Information
PRC	Precision Recall Curve
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SVD	Single Value Decomposition
T2V	Tweet to Vector
TF	Term Frequency

1. INTRODUCTION

Modern biosurveillance systems rely upon a variety of electronic signals to detect potential disease outbreaks. Typically, data from hospital emergency department admissions, emergency response telemetry, and various social network media are analyzed to identify indications of potential disease processes. Two categories of algorithms are vital to consistent and timely detection of disease from electronic data streams: anomaly detection and data fusion (Hopkins et al. 2017). Anomaly detection algorithms identify outliers in electronic signals that are atypical and might represent disease events. Data fusion algorithms allow the integration of multiple contemporaneous data feeds to generate a trigger signal that may not have been measureable on single channel data streams.

This report examines a range of novel anomaly detection and data fusion algorithms with potential to serve as biosurveillance anomaly detection and data fusion components. The evaluated algorithms are representative of a class of processes known as biologically inspired algorithms, meaning that the algorithms mimic natural biological process in some way.

For brevity, this report summarizes prior published work resulting from a recently conducted Laboratory Directed Research and Development study. Design details on the algorithms mentioned on this report and results of performance testing can be found in a number of reports and

2. BIOSURVEILLANCE DATA

Large scale biosurveillance systems rely on information derived from many sources to infer an incipient disease outbreak or bioweapons attack. Typical data sources for biosurveillance include hospital emergency department (ED) records, pharmaceutical prescribing and purchase information, employee absenteeism information from large employers, and diagnostic laboratory results. Each of these data sets may contain personally identifiable information (PII) and, thus are not publically available. Production biosurveillance systems execute detailed data use agreements (DUA) with data providers to ensure that PII is not accessible, and that strong data security procedures are in place. Researchers often find it difficult to identify and acquire useful datasets for development and testing of improved anomaly detection and data fusion algorithms.

For this study deidentified public health data have been obtained from a variety of sources. Emergency department datasets were obtained through data-sharing agreements with public health departments in North Carolina, Massechusets, and New Mexico. Summary influenza data was obtained from US Outpatient Influenza-like Illness Surveillance Network (ILINet) compiled by the US Centers for Disease Control and Prevention (CDC).

Biosurveillance anomaly detection and data fusion algorithm research requres varied data sets to be used for development and testing to ensure that the system has been adequately tested and calibrated to perform in conformance with requirements. Initial model development and testing has largely relied upon synthetic data sets. Data sets generated using methods described by Levin and Finley (2016) and Levin et al. (2018) permit the development team to rapidly determine performance characteristics of system components by testing against synthetic data sets specifically configured to exercise the capability or feature being developed or refined. To support integrated system testing under controlled conditions for noise and anomalous signals, coherent synthetic data sets across all data feeds have been used as software components are integrated into subsystems. In addition, synthetic data permits performance evaluation for rare events such as overlapping outbreak events that would not be expected in typical training data.

A variety of electronic health record data feeds and syndromic surveillance data sets have been incorporated into the testing and evaluation processes once components are performing well on purpose-designed synthetic data. Sandia's active data use agreements with a variety of entities, including NC-DETECT, Boston Health and the New Mexico Department of Health, ensured the availability of comprehensive data sets covering diverse geographical settings and seasonal variabilities for many target diseases. These multimodal data sets have provided a rich variety of well-studied natural outbreak signatures upon which to refine fusion and anomaly detection capabilities.

3. IMMUNE SYSTEM ANOMALY DETECTION AND DATA FUSION

National scale biosurveillance systems monitor a wide range of electronic signals to identify early evidence of possible disease outbreaks or bio-weapon attacks. Early detection of an emergent disease outbreak is crucial for a timely and cost-effective response. Signs of emergent outbreaks can be hidden inside high-dimensional data that is both noisy and incomplete. Biosurveillance detection of these events requires novel data analysis and classification techniques. The design and implementation of such detectors is an ongoing task in the field of electronic biosurveillance (Shmueli and Fieinberg 2006; Unkel et al. 2012; Gajewski et al, 2014). Current detection mechanisms often derive from established methods of time series analyses from other domains (Goldenberg et al. 2002; Cheng et al 2012; Schmueli 2013) and may not be best suited to deal with the complexity of modern biosurveillance data.

Current methods applied to health record surveillance rely on standard statistical approaches such as control chart algorithms (Morton et al. 2001; Woodall et al. 2006) and Bayesian Belief Networks (Burkom et al. 2011). While these methods perform well on specific types of data feeds, they often don't handle multivariate data (control charts), continuous data (Bayesian Belief Networks), and frequently do not scale well to the larger data sets available for biosurveillance.

To address these limitations, Levin and Finley (2017) turned to a known natural distributed anomaly detector. The adaptive immune system is able to maintain a distributed repertoire of lymphocytes that can recognize and respond to foreign pathogens while avoiding any response to healthy tissue. Applying naturally inspired algorithms in new domains is an established practice. Previous work using immune-inspired classification approaches have been successfully used to detect fraudulent ATM transactions (Ayara et al. 2005), unauthorized intrusions Greensmith et al. 2006), anomalous port scans (Greensmith and Aiklelin 2008; Greensmith et al. 2010), and invalid online media streaming purchases (Huang et al. 2010).

3.1. Dendritic Cell Algorithm

Levin and Finley (2017) initially evaluated the Dendritic Cell Algorithm for potential applicability to biosurveillance anomaly detection. They implemented the algorithm and tested it against challenging synthetic datasets to evaluate its performance relative to traditional statistical based anomaly detectors.

3.1.1. Algorithm Description

The innate immune system is adept at recognizing and responding to a wide variety of foreign pathogens. The dendritic cell algorithm was by Danger Theory, first described by Matzinger (1994). Ubiquitous dendritic cells continuously ingest free antigens in tissue, thus maintaining individual repertoires of their local environment. These dendritic cells simultaneously sample the local molecular profile. In the event of an infection, epithelial cells will secrete specific 'danger' molecules. Upon detection of these danger signals, dendritic cells will classify their entire current repertoire of ingested antigens as foreign and travel to local lymph nodes to present these particles to lymphocytes to initiate an immune response.

The process of classification through the combination of local sampling and exogenous danger signals has been implemented as the Dendritic Cell Algorithm (DCA) by Greensmith et al. (2006). The DCA classifies data by first mapping input data (anti- gen) to a combination of three values: safe, danger, and PAMP (pathogen associated molecular pattern) signals. Simulated immature dendritic cells sample and aggregate these input values using a static trans- formation matrix for a limited amount of time, after which the dendritic cell transitions to a mature (infection) or semi-mature (no infection) state depending on the sampled data. Upon transitioning, all sampled data of the dendritic cell are given a token indicating either mature or semi-mature. After the sampling phase has run to completion, each individual data point is classified as safe or dangerous based on a comparison of the relative token counts to a given threshold value.

3.1.2. Experimental

The DCA was implemented and tested with representative biosurveillance data sets. Results were not encouraging, with the algorithm identifying strong potential disease signals, but not detecting more subtle signals in noisier datasets (Figure 1). In addition to the algorithm's detection performance, it proved to be very difficult to tune effectively, since the parameters of the algorithm do not correspond to quantities of relevance to the biosurveillance domain. Finally, the DCA's performance was judged inadequate since its design requires that raw input first be classified with a traditional anomaly detection routine, such as CUSUM. Thus the DCA cannot produce anomaly detection superior to the flawed control chart methods currently in use. Further research on using the DCA for biosurveillance anomaly detection was not pursued (Levin and Finley 2017).

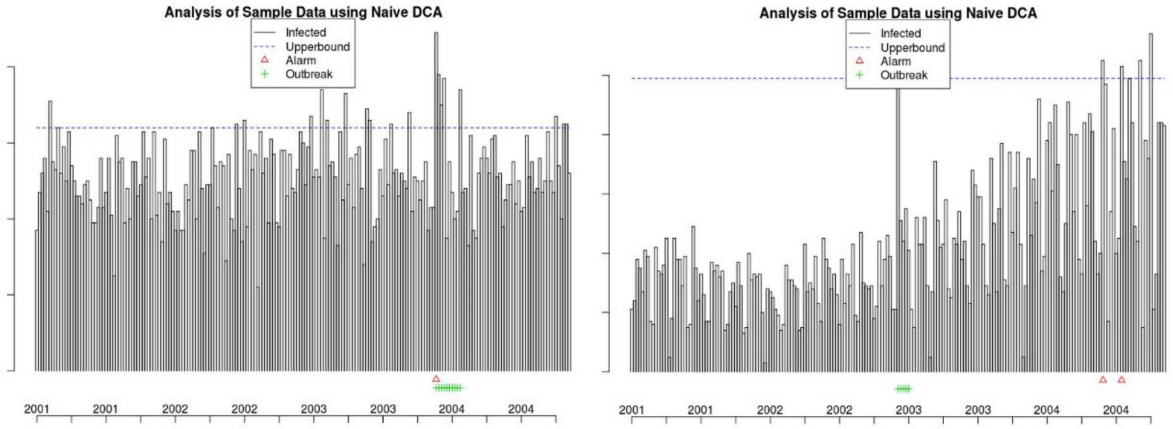

Figure 1. Naive DCA Implementation. Without exogenous transformation of the input signal, the DCA is unable to recognize non-stationary data. The DCA was tested against a stationary (left) and non-stationary (right) data set, each with one simulated outbreak (green +). The DCA is able to properly identify the outbreak in the stationary data set (red triangle), but cannot separate the outbreak from the high baseline values in the non-stationary data set.

3.2. Negative Selection

The innate immune response is able to sample local antigen and respond to outbreaks as directed by an exogenous danger signal, but is not able to learn or adapt to items previously seen. The adaptive immune system is responsible for maintaining a repertoire of lymphocytes that can recognize and

respond to pathogens previously classified as anomalous by the innate immune system, while avoiding any response to healthy self antigens. The biological mechanism that generates and filters the T cell population is known as Negative Selection (NS) (Nossal 1994). T cells that survive the NS process should not be able to bind to any self molecules and therefore anything they do bind to can generally be considered foreign.

3.2.1. Algorithm Description

Levin and Finley (2017) implemented a mechanistic version of the biological NS algorithm for use as a biosurveillance detector. The artificial NS classifier handles multi- dimensional data time series data. Artificial T cells have a independent detector for each data dimension and will recognize a data point if each of its detectors react with each component of the sampled data point. Each input dimension can be one of three possible data types: count, indicator or categorical.

To create a complete T cell repertoire, T cells are generated randomly such that they contain unique values and specificity in each dimension. To generate random T cells, first a training data set with no known disease outbreaks is examined. T cell values are chosen uniformly from the upper and lower ranges of numerical data, and sampled without replacement from categorical data. Indicator specificity is constrained to be within a min and max numerical range of the generated center value and categorical speci- ficity is set as the size of the T cell's subset of possible items. The minimum size constraint on ranges corresponds to the biological mechanism known as Posi- tive Selection, and ensures that generated T cells are appropriately general .

Once generated, each T cell is tested against the selected training set of baseline data. If a T cell reacts to any baseline point, the T cell is removed from the population. This process is basis for the Negative Selection designation and ensures that the remaining T cell repertoire does not react with any data known to be quiescent. Future data that react with any of the surviving T cells will be classified as anomalous. Because the remaining T cells survived both positive and negative selection, they can be considered both appropriately specific and general.

3.2.2. Experimental

The anomaly detection performance of the NS algorithm was evaluated against a number of different data sets. The algorithm performed well against all tested data sets, delivering anomaly detection capabilities equaling or exceeding that of traditional statistics based anomaly detectors (Figure 2).

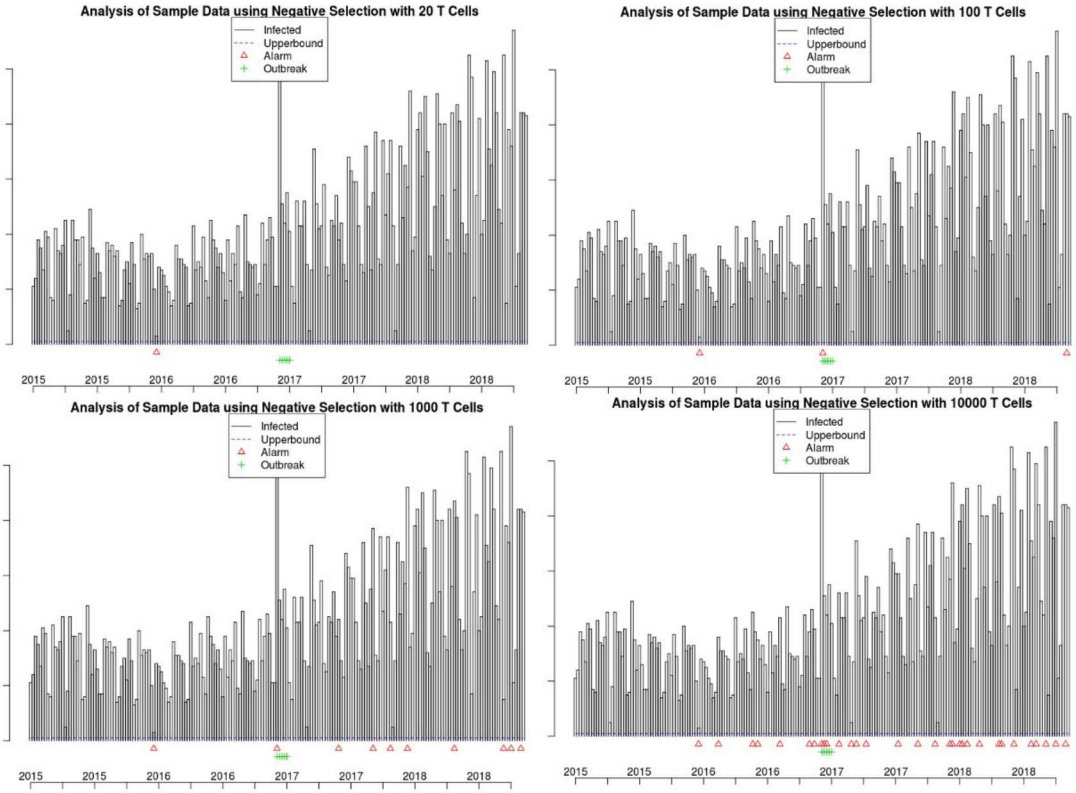

Figure 2. Negative selection anomaly detector sensitivity and specificity. Shows how the number of T cells affects a Negative Selection classifier on a non-stationary data with a single outbreak (green +). Too few T cells (upper-left) results in poor coverage of the anomalous space and an inability to detect outbreaks (red triangles). Too many T cells (lower panels) results in over-saturated coverage and an overabundance of false positive classifications.

3.3. Negative Selection Data Fusion

Levin et al. (2018) investigated the capacity of the Negative Selection (NS) algorithm to simultaneously process multiple channels of input data. In nature, immune systems sense large numbers of antigens simultaneously and trigger cascade responses to mount an effective response. This inherent property of native parallelism in functioning is particularly intriguing from the standpoint of alternative algorithms for biosurveillance. A recent review of outstanding research issues in biosurveillance cited the need for vastly improved data fusion approaches to enable emerging disease outbreakks to be detected from faint signals in multiple channels, where each individual signal may not be strong enough by itself to rise above the detection threshold, but complementary signals recorded across multiple channels can denote a signal of interest when considered as a whole (Hopkins et al. 2017). Additionally, ensembles of NS detectors are not only capable of multi-channel anomaly detection, but the data channels may represent different data types resulting in multimodal data fusion capabilities.

3.3.1. *Experimental*

To explore efficient anomaly detection across multiple channels, Levin et al. (2018) enhanced the NS biosurveillance detector described by Levin and Finley (2017) by incorporating bagging and a parallel implementation model that permitted more efficient computation by allowing efficient distribution of training and testing loads to be parked on individual cores of a multicore CPU architecture. Effective parallelism of the improved architecture was tested using a nine-dimensional synthetic data set consisting of 100 individual generated one-year sets each with a simulated anthrax incident for the detector to predict.

By enhancing the basic NS algorithm to better exploit the native distributed nature of the immune system approach, Levin et al. (2018) effectively demonstrated the capability of T-Cell detectors to operate efficiently on high dimensional data with a high level of sensitivity and specificity (Figure 3).

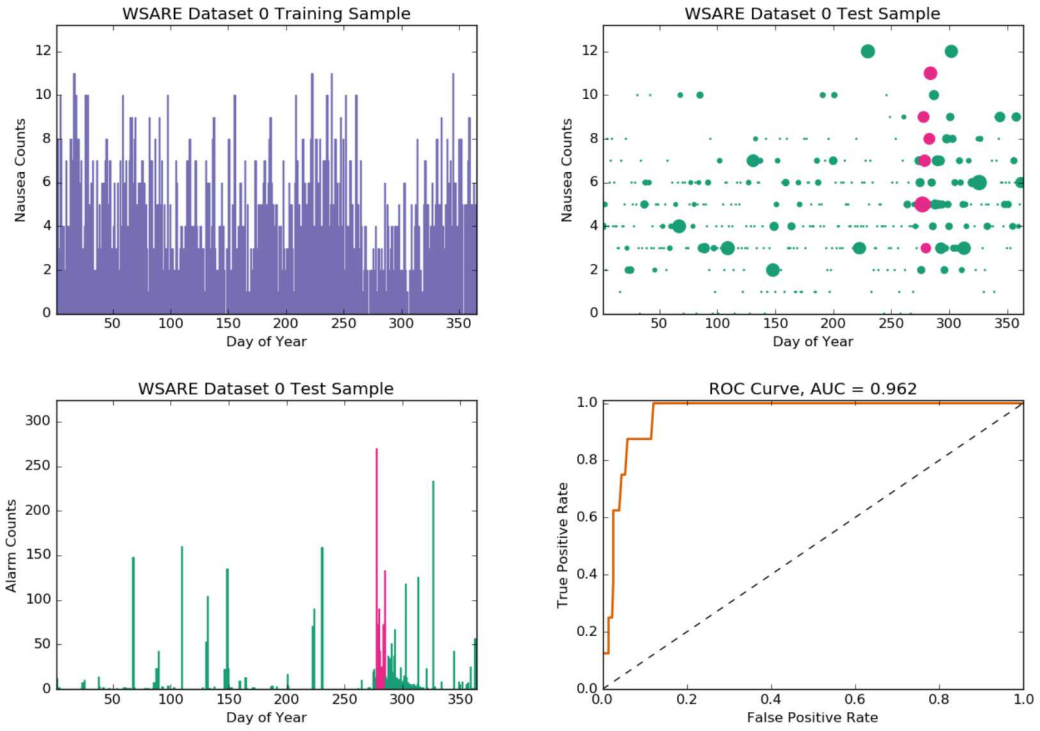

Figure 3. Application of negative selection for anomaly detection on a sample dataset. Upper Left: Nausea counts of the WSARE dataset as an example dimension. Upper Right: The generated detectors were applied to the second year of the dataset. A true anomalous outbreak occurs at day 277 and is shaded pink. The size of each data point represents the number of detectors that overlap the point. Lower Left: The alarm rate for each day of the test set, analogous to the size of the data points in the upper right plot. Right: The ROC curve for the generated alarm rate as compared to the true outbreak.

3.4. Biosurveillance Network Topology

Flanagan et al. (2018) discussed alternative topologies for biosurveillance information processing networks based on two well-understood biological analogs. Both ant colonies and adaptive immune systems function through distributed search and processing paradigms. The non-centralized processing and sharing of information enables both archetypal systems to respond more quickly to local events, and to effectively segregate spatiotemporal regions of disturbance from the unaffected portion of the network, ensuring that the gross behavior of the network is largely maintained while localized issues are isolated and resolved efficiently. Applying this motif to biosurveillance, they postulated that shorter distance information flow, and smaller effective areas of decision effect could similarly promote a more efficient national biosurveillance system, where nascient outbreaks are detected and addressed locally, before they can spread long distances. Focusing on immune systems, Flanagan, et al. proposed that the T-Cell-based biosurveillance anomaly detection and data fusion concepts described by Levin and Finley (2007) and Levin et al. (2008) could scale to a corresponding conceptualization of a regionalized biosurveillance network system patterned after the lymphatic system. In the lymphatic system, T-Cell signals from adjacent tissues are monitored, and additional adaptive cells are dispatched to the points of inflammation from which the T-Cell signals emenated.

Extending the analogy of Flanagan et al (2018) to the US biosurveillance system, two end member cases can be considered (Figure 4). In the centralized topology, information processing and decision processing are focused at a single location, such as at the CDC in Atlanta, GA. Alternatively, information processing and interpretation could be distributed throughout the country, specifically at local, county, or state health departments, thus speeding and simplifying response to locally-derived disease outbreaks. Operationally, the national biosurveillance network functions as neither end member example, but as some hybrid model, with budgets and available expertise often dictating whether local or centralized resources are used. Flanagan's contribution to the discussion is to cite numerous examples from natural systems showing that biosurveillance systems tending toward the distributed pole could be expected to be more responsive to local needs and more cabable of rapid response than centralized topologies which would presumably be more automated and thus less expensive to provision and operate.

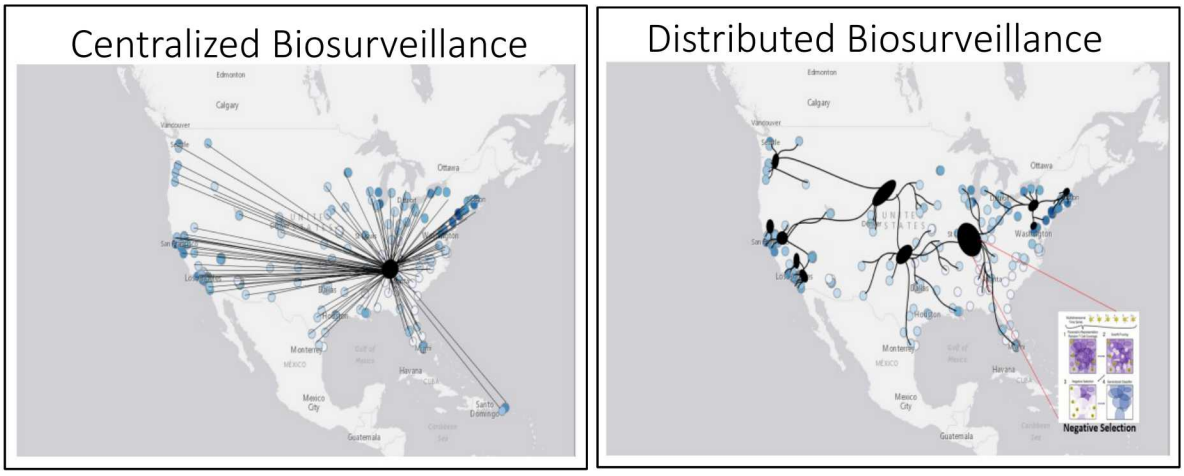

Figure 4. Conceptual models of centralized and distributed national biosurveillance networks.

Beyeler and Finley (2018) presented a mathematical framework to quantify the relative advantages of distributed and centralized biosurveillance network topologies. The framework examined the relative importance of speed of detection and efficiency of communication, which favor centralized topologies versus quick respone and use of local expert knowledge to reduce false-positive detection. By considering the virulence of the pathogen and also potential for a disease propagation pathway to exploit air travel to rapidly spread, the mathematical framework enabled biosurveillance network topologies to be approached from a pathogen-specific standpoint, thereby suggesting alternative topologies that are optimized to protect public health or provide bio-weapon attack alerts for specific classes of disease causing agents. Initial model runs on synthetic data reported by Beyeler and Finley (2018) should be followed up shortly by model runs on curated natural biosurveillance data, thus enabling a more rigorous evaluation of the biologically-based concepts proposed by Flanagan et al. (2018).

4. NEURAL NETWORK ANOMALY DETECTION AND DATA FUSION

Concurrently monitoring multiple data sources to derive reliable alerts for potential disease outbreaks or CB events can make biosurveillance much more accurate and sensitive. However disparate biosurveillance data streams often differ in value type, update frequency, signal-to-noise ratio and timeliness. Both the value and the difficulty of generating unambiguous event alerts from noisy, disparate data streams is anticipated to grow in the near future as the number, complexity and size of new online data sources will increase markedly.

Biosurveillance detection systems monitor a wide range of near-real-time data feeds to identify indicators of disease outbreak or CB activities. These feeds can include social media posts or searches, hospital admission data, meteorological conditions, population density and movement, environmental sensors, laboratory test results, and public health reports among others. Inherent in the collection and processing of multiple concurrent data feeds is the assumption that diverse sources can provide a more reliable indication of a true outbreak and support the resolution of ambiguous contradictory indicators that could result in false detection. Thus, the proposed research and development effort defines the problem to be the generation of defensible alerts based on the combined information from disparate data flows while concurrently minimizing possible false indications.

4.1. Traditional Methods

Biosurveillance systems currently in use apply traditional statistical methods to identify anomalous data values that could indicate a disease outbreak or bioweapon attack. These statistical methods are well understood and provide consistent performance across different data types. Statistical data fusion methods have been explored for for use in biosurveillance, but to date, few production biosurveillance systems incorporate data fusion capabilities.

4.1.1. Anomaly Detection

Biosurveillance systems rely upon algorithms to recognize time-series data points whose values are anomalously elevated relative to preceding data points; it is often interpreted as indicating that a disease outbreak in occurring. Anomaly detection algorithms in current use vary in complexity from simple sliding-window running-average calculations to sophisticated time-series statistical and machine learning approaches. These conventional statistical multivariate anomaly detectors have two primary limitations: (1) they encode temporal patterns via auto-regressive and seasonal methods, which are not very useful if temporal patterns of outbreaks change from year to year; and (2) they try to relate various data streams via Gaussian models, which are not always appropriate (e.g., in case of low counts, where Poisson distributions are preferred).

A few statistical biosurveillance anomaly detectors have attained relative prominence. Cumulative Sum statistics (CUSUM) and Exponential Weighted Moving Average (EWMA) are the most common biosurveillance anomaly detection methods in current use (Shmueli and Burkom, 2010). These methods are widely available in the Surveillance R package (Höhle, 2007). CUSUM is a control chart method that displays the cumulative sums of the deviations of each sample value from the anticipated target value. EWMA differs from other control chart methods in that it gives less weight to data points that are further removed in time. CUSUM has been shown to be particularly useful in error detection for data sets characterized by large changes over an extended period of time (Han et al., 2010), or for anomalies exceeding one standard deviation unit (de Vargas et al. (2004))

EWMA, on the other hand, was better at detecting anomalies not exceeding 0.5 standard deviation units, especially if the occur early in the time series. CUSUM and EWMA will be compared with our RNN-based anomaly detector when we assess its performance.

4.1.2. Data Fusion

Electronic biosurveillance systems have traditionally relied upon standard statistical data fusion methods to integrate data streams. Multivariate control charts in various forms are often applied to multiple biosurveillance data streams (e.g. Fricker, 2007). Multivariate space-time clustering has shown the capacity to incorporate spatial and temporal inhomogeneity in multivariate biosurveillance data (Burkom, 2003; Kulldorff et al., 2007). Lau et al. (2012) demonstrated that dynamic linear models can improve situational awareness of seasonal influenza activity from multiple data streams.

More sophisticated fusion approaches specifically tailored to biosurveillance have been reported but have yet to be incorporated into production systems including, for example, Bayesian networks (Mnatsakanyan et al., 2009; Burkom et al., 2011). Corberán-Vallet (2012) reported improved consistency among incidence-count data sets by sophisticated Bayesian statistical models. Multivariate analysis based on generalized branching process methods were reported by Paul et al. (2008) to achieve improved results for influenza and meningococcal disease in Germany. Schiöler and Frisen (2012) derived a maximum-likelihood estimator to underpin a generalized-likelihood multivariate detection method which was shown to be robust across multiple data streams incorporating spatial variability for influenza outbreaks in Sweden. Ray et al. (2012) demonstrated that Bayesian fusion of disease case counts together with spatial and temporal information significantly improved the sensitivity of outbreak anomaly detection compared to depending solely on case counts.

4.2. Biosurveillance Challenges

The existing statistical methods cited above have been shown to generate indicators of disease activity from isolated multivariate biosurveillance data sets. However, existing methods lack important features needed to effectively extract consistent actionable outbreak signals across the expected range of potential input data streams. The factors we assert to be required for a general-purpose production-level biosurveillance data fusion system include:

- Utilization of multiple data types. Ideally the system should incorporate integer count data, categorical data, and real-valued sensor data.
- Resolution of temporal lags in correlated signals between data streams
- Resolution of missing or aperiodic data within time-series streams
- Recognition of historical patterns in multivariate data streams that have been associated with prior outbreaks, and utilization of these patterns in weighting potential alert statuses
- Incorporation of maximal information content in the data streams, which has historically been precluded by the existing options of early or late fusion.

4.2.1. Multiple data types

While simple case-count data have traditionally served as the basis for biosurveillance outbreak detection, modern systems should a wide range of data types for analysis. Fusing multiple data streams of the same basic type (e.g. integer case counts) can be accomplished with standard

methods. Consistent integration of streams of fundamentally different types (e.g. real-value and categorical) into an alert assignment weighting is not well-handled by existing methods

4.2.2. Temporal lags

Different data streams exhibit distinct temporal signatures for correlated events. For example, Ray and Brownstein (2013) demonstrated that the time-series response of HealthMap news item counts has a different characteristic response profile than the epi curve of the underlying disease incidence. The news-based HealthMap feed rises very steeply and peaks while the corresponding case count data are still building. As the news value of the potential outbreak wanes, the HealthMap activity level drops rapidly, while the incidence counts are still rising toward a peak. Ray and Brownstein implemented sophisticated statistical time-series smoothing procedures to overcome this particular issue, but similar systematic lags among different data feeds could presumably lead to the assignment of multiple offset peaks to separate stimuli rather than to a single event inducing different characteristic temporal signatures along different data channels.

4.2.3. Missing data

Standard biosurveillance fusion methods, as well as those used in other domains such as battlefield situational awareness, often rely upon regular coordinated updates from multiple data feeds, often at hourly or daily intervals. However, electronic health records (EHR), hospital admission data, and diagnostic laboratory results are often posted as they become available, leading to irregular data frequency with substantial and unpredictable gaps between reported readings. Effectively merging information from multiple irregularly-updated or aperiodic streams with strictly periodic update data feeds such as hourly temperature or absolute humidity feeds is analytically challenging.

4.2.4. Pattern Recognition

Traditional biosurveillance data fusion methods often rely upon comparison of a composite value to fixed thresholds to recognize an anomaly leading to the issuance of an outbreak alert. This naive approach does not make use of information gleaned from prior outbreak incidents which could indicate complex patterns among the monitoring data feeds prior to or coincident with the outbreak. This inability to learn from prior experience prevents detection of both seasonal disease patterns and signatures of the more rare historical outbreaks using present methods.

4.2.5. Fusion Order

Methods used to integrate data streams can be grouped into two categories, *early fusion* or *late fusion*. Each of these approaches can have substantial advantages and drawbacks.

In early fusion approaches, disparate data feeds are amalgamated into a single indicator and the subsequent anomaly detection determination is applied to that combined value. This approach enables the data streams to be combined using methods that are optimized to exploit synergistic properties of the incident data streams. However, since anomaly detection operates on the combined data feed, early fusion methods can preclude choosing analytical processes specifically tailored to extract maximal information from each separate data feed.

In late fusion, data feeds are analyzed for anomalies individually. Anomalies noted through extensive analysis of single feeds are then evaluated at the fusion operation through, for example, logical AND or OR operations to determine if the multivariate input values require that an outbreak alert be issued. The advantage of late fusion is that analytical algorithms can be specifically tuned to best

handle individual data feed characteristics. The major drawback is that subtle correlated indicators which may exist among data feeds cannot be determined.

4.2.6. Deep Learning Alternatives

Recent research in AI has produced sophisticated mathematical methods with the potential to address the outstanding issues in data fusion for biosurveillance noted above. Specifically, multimodal deep learning (MDL), the innovative state-of-the art approach to multivariate data fusion, has the potential to greatly improve sensitivity, specificity and timeliness of outbreak detection from multiple biosurveillance data streams. This section provides descriptions of technologies behind MDL, and describes examples of applications of constituent methods to difficult analytical problems.

In contrast to the traditional statistical data fusion methods discussed previously, machine learning (ML) methods enable computers to learn from data and make data-driven decisions and predictions without being specifically programmed to address those particular decisions. Machine learning systems are trained by exposing them to large quantities of data. From these training data, ML systems can learn to discriminate patterns and relationships that are often quite subtle. Once trained, ML systems can evaluate new data based on learned relationships and patterns to rapidly categorize the new data or to predict missing values.

Recently a family of machine learning methods known collectively as Deep Learning (DL) have shown remarkable ability to detect and operate upon subtle patterns and relationships within complex data. Deep learning systems have revolutionized image processing, language translation, autonomous vehicle guidance and many other fields over the past three years. While DL has not to date been applied to biosurveillance data fusion, unique features of DL systems designed for image processing and natural language processing can be repurposed to provide new and powerful capabilities for the generation of improved outbreak detection alerts from multivariate biosurveillance data feeds.

Deep learning relies upon Artificial Neural Networks (ANN), software structures that abstractly mimic nerve cells and their connections. While ANNs have been known and used for decades, recent advances in computational power and the availability of large training data sets have catalyzed development of very large networks capable of encoding complex relationships and patterns.

4.2.7. Sequence Analysis

Many of the best known advances in Deep Learning have involved discovery and categorization of patterns in images and videos using an architecture known as convolutional neural networks or CNNs. While CNN's are flexible and able to be applied to many different problem domains, they are not well-suited to biosurveillance data fusion due to their reliance on fixed size inputs. Biosurveillance data streams are typically time series and are, by nature, of variable length. Recurrent neural networks (RNNs) are particularly well-suited to problem domains characterized by input data of indefinite size such as language. RNNs have recently been applied to language translation, lip reading, and speech-to-text conversion (e.g. Socher et al., 2014). While the Natural Language Processing (NLP) tasks addressed by RNNs may eventually have application to biosurveillance, it is RNNs' ability to detect patterns in time-series data that is most relevant to biosurveillance applications.

Although RNNs have been applied in a variety of challenging pattern recognition and time-series analyses, numerical instabilities can limit their utility for real-time mission-critical applications. RNNs use feedback loops within their architecture to recursively improve the quality of weights calculated through the iterative optimization step. The recursive processing that characterizes this architecture can lead to vanishing gradients or exploding gradients within the network optimization processes, preventing convergence. Additionally, the recursive nature of RNNs limits the ability of these networks to maintain long term memory of prior states to at most two update cycles, decreasing utility for many time-series tasks working to identify patterns in data based on examples that were previously encountered.

4.2.8. Time-series Analysis

Extensions to basic RNN topologies have produced a powerful tool for mission-critical time-series analysis. Long Short-Term Memory (LSTM) networks are a modification to traditional RNNs which simultaneously resolve numerical instabilities while greatly extending the ability of the network to incorporate long-term dependencies into pattern recognition tasks (Hochreiter and Schmidhuber, 1997). LSTMs have proven very useful to a wide variety of tasks that require processing sequences of data. Vohra et al. (2015) use deep belief networks coupled with LSTMs to analyze musical sequences. Zhao et al. (2016) applied LSTM networks to monitor the health of electrical motors in factory settings and reported cases where the monitoring permitted timely replacement prior to failure. LSTMs were trained on the actions of first responders during crises and provided timely guidance to method selection for unfolding events (Nguyen et al., 2016). Applying LSTMs to electronic health records allowed Choi et al. (2016) to reliably predict the propensity for heart failure in patients with far greater precision than standard statistical approaches. Brownlee (2016) demonstrated that LSTM analyses can simultaneously predict observed periodicity and non-stationarity in airline passenger census data in spite of the limitation of very small training data sets. As this brief discussion of LSTM-enabled analyses demonstrates, deep learning-based time-series techniques have rapidly eclipsed traditional statistical approaches for investigation of challenging data sets.

4.2.9. Anomaly Detection

The power of deep learning methods, particularly RNN-LSTM network analysis, has been successfully applied to anomaly detection (AD) across a range of domains. LSTM-based AD approaches are characterized by improved performance in noisy data settings and the ability to resolve faint anomalies. George and Huerta (2017) applied deep learning methods to detect subtle signatures of black-hole collisions in electromagnetic cosmic background noise. This deep-learning-enabled anomaly detection example is particularly noteworthy in that the authors used neural networks to both detect the anomaly and to classify the signal regarding the probable mass of the black hole pair emitting the signal. Malhotra et al. (2015) exploited the unique capability of LSTM networks to concurrently capture both long-term and short-term periodicities in complex data streams, periodicities that enabled the reliable detection of previously undocumented anomalies in reference time-series data sets from space shuttle telemetry, electrocardiograms, and engine sensors. An extensive study of LSTM-based anomaly detection by al Dosari (2016) demonstrated conclusively that unlike traditional AD methods, LSTM can detect not only anomalous point values, but also anomalous complex patterns in time series data.

4.2.10. Data fusion

Recent work on deep learning data fusion approaches have shown the promise of deep learning methods for integrating disparate data sets. Foundational work by Ngiam et al. (2011) clearly demonstrated the potential gain in information content from use of deep learning methods for multivariate data streams. Using audio and video data streams to automate lip reading, Ngiam et al. showed that while training a neural network with multimodal information provides rapid, consistent classification of speech patterns, this novel method also enables reliable classification performance when only one data feed was available. Wu et al. (2015) similarly showed that data fusion systems composed of multiple deep neural networks consistently outperformed traditional canonical correlation fusion methods on video and audio feeds. Ma et al. (2015) showed how multimodal deep learning neural networks can reliably categorize images and their representative captions. Akita et al. (2016) use deep LSTM networks to fuse financial time series data consisting of numerical and textual information.

4.3. Applying Deep Learning to Biosurveillance

As our discussion of relevant literature shows, multimodal deep learning (MDL) approaches to time-series analysis, anomaly detection, and data fusion have rapidly exhibited the potential to eclipse statistical-based approaches in a wide variety of domains. Additionally, MDL approaches can directly address identified outstanding issues with biosurveillance data fusion.

4.3.1. Missing and Aperiodic Data

While traditional biosurveillance data fusion methods struggle with missing data and data feeds with differing granularity, recent studies demonstrate that MDL approaches have the potential to address this biosurveillance issue directly. Lipton et al. (2016) adapted standard LSTM architectures to enable adaptive data imputation for multivariate time-series electronic health record data feeds to resolve significant gaps in individual inputs. Che et al. (2016) reported on architectural modifications to standard LSTM-type networks to specifically optimize performance in patient records with significant missing data.

4.3.2. Early vs Late Fusion

A recognized limitation of traditional data fusion methods is that either of the fundamental approaches, early fusion or late fusion, imposes limits on the eventual quality of the combined data product. Recent advances in MDL data fusion methods have demonstrated the potential advantages of hybrid fusion approaches which overcome the recognized limitations of either early or late fusion. Gandhi et al. (2016) devised a unique MDL architecture with independent pathways which apply early fusion to temporal data streams, and late fusion to non-temporal values resulting in greatly improved performance in image captioning processing.

4.3.3. Correlated Temporal Lags

Indicators of potential outbreaks can occur at different times in individual biosurveillance data streams. Statistical methods struggle with correlating these lagged signals and assigning them to a single event. MDL data fusion methods can outperform conventional methods on multivariate temporal lags. Supervised training of LSTM networks with specific historical examples where lagged indicators exist can produce default classification behavior that routinely groups these lagged signals into a single event. While this feature of the MDL data fusion approach has not been reported in the

literature, empirical tests to establish its feasibility are straightforward given a representative suite of biosurveillance data.

4.3.4. Pattern Recognition

While traditional anomaly detection methods are largely limited to recognizing novel outbreaks, MDL methods can concurrently recognize both the signatures for novel outbreaks and specific signatures for previously encountered outbreaks. MDL-based anomaly detection operates on a neural network trained by exposure to previously recorded data. By feeding the neural network labeled historical data in which example outbreaks are noted, the system learns to detect distinctive patterns in incoming data in addition to deviations of incoming signals from a threshold. This constitutes a principal potential advantage of MDL-based anomaly detection over traditional statistical methods.

4.3.5. Summary

Multimodal deep learning methods exhibit tremendous potential to improve routine multivariate monitoring of electronic data for indications of disease outbreak or CB events. Deep learning AI approaches have revolutionized natural language processing, time-series data analysis and multivariate data fusion. The powerful capabilities developed in these divergent domains can be leveraged to provide similar improvements in routine biosurveillance monitoring.

4.4. Experimental Evaluation of Deep Learning Approaches

A prototype biosurveillance data fusion application based on Multimodal Deep Learning (MDL) was implemented and tested. The application monitors multiple electronic data streams and generates an alert when patterns within the data streams indicate likelihood of a disease outbreak or CB event. Data streams monitored by the application can be of various types, update rates and signal-to-noise ratios. The application was built upon a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture, capable of learning characteristic patterns in multimodal biosurveillance data streams which indicate anomalies of interest. Initial implementation has focused on fusion and anomaly detection within simple pairs of data streams (e.g. integer case counts and real-valued sensor feeds). Rigorous integrated validation of the fusion and anomaly detection processes to ensure that system performance measured by standard metrics is maintained as additional, and more complex, synthetic data streams are incorporated has not yet been performed. Further testing and tuning of the MDL data fusion system incorporating biosurveillance data could ensure that design goals and performance benchmarks are maintained with progressively more challenging field data feeds. Eventually, standardized biosurveillance data streams should be evaluated with the prototype application, with performance metrics evaluated to ensure required performance on production level data. Once acceptable performance is demonstrated on the system using data feeds representative of national biosurveillance data feeds, a prototype port of the application capable of real-time analysis on real-time biosurveillance data feeds could be be implemented as further proof of concept.

4.4.1. Data Fusion Approach

MDL data fusion systems have been demonstrated to provide robust and efficient integration of disparate data types for a range of domains. However, application of this state-of-the-art data fusion method to biosurveillance data feeds has not been reported in the literature. Thus adaptation of published MDL data fusion methods to the unique characteristics and limitations of biosurveillance

data constitutes a significant research and development component of SNL's biosurveillance research to date

Conceptually, the MDL data fusion system uses LSTM-RNNs to combine multiple concurrent data feeds into a single signal or decision variable that can then be mapped to the likelihood of a disease outbreak or CB event. As with many machine-learning-based analytical processes, use of these LSTM-RNNs for data fusion incorporates two distinct phases: training and classifying. For biosurveillance data fusion, training entails feeding large historical archives of the data feeds to the application before using the networks to perform actual data fusion. During the training operation, the networks calculate specific numerical weights and non-linear transformations that encode variational patterns observed within the training data set. For the biosurveillance data fusion training task, supervised learning approaches were applied, labeling each item in the historical time series data feeds as to whether they occurred during an active disease outbreak or CB event or not. This labeling of historical example data feeds enables the neural network to differentiate patterns in the combined data sets that indicate disease outbreak from patterns that do not. Once training was complete, the neural network eventually was applied to classify synthetic real-time composite data feeds to determine whether the unlabled inputs more closely match the learned features indicative of an active disease outbreak or the features indicative of a non-outbreak condition.

4.5. Experimental

Two variants of MDL data fusion approaches demonstrate the potential of deep learning methods for resolving biosurveillance issues common with single-channel decision support, or traditional statistical methods of data fusion

4.5.1. Resolving Temporal Offsets

A prototype of an MDL data fusion application is presented to demonstrate the capabilities of this innovative approach to generate meaningful outbreak alerts from multiple data streams. The structure of the example LSTM-RNN neural network application is shown in the schematic illustration in Figure 5. The example system is composed of four layers: a three node input layer, a five-node densely connected LSTM layer a 5 node hidden layer and a single node sigmoid output layer. The decision threshold of the final output layer was set at 0.5 to provide unbiased prediction between two classes. The network was trained using the Adam optimizer with a learning rate of 0.001 and a binary cross-entropy loss function. The network used a batch size of 100 and was trained for 100 epochs for this example.

Two multivariate synthetic data sets were constructed for this example, each consisting of three separate inputs: clinic visits, news report counts, and day of the year. The system was trained for 10,000 days of synthetic data with known outbreak periods. A test data set consisting of 1,000 days of clinic visits and news-report counts was constructed to test how well the neural network predicts an outbreak and fires outbreak alerts.

Results of the example data fusion run are shown in Figure 6. Notice that news report counts and clinic visits exhibit a temporal lag as described by Ray and Brownstein (2013) where the news counts peak and decay before the clinic visits increase. The neural network accurately predicts two outbreaks, (Day 350-400 and 600-680). Notice that temporal lag between the data peaks is not incorrectly interpreted to represent two separate outbreaks. Similarly, the peaks that occur in the

daily clinic visit data set at around Day 100 and Day 780 do not trigger an alert because a correlated peak in news report counts was not present. This trivial example demonstrates that MDL data fusion systems exhibit reasonable behavior on test datasets, and easily handle the temporal lag issue which challenges traditional statistical data fusion algorithms.

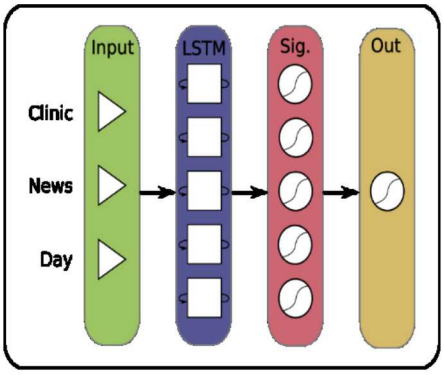

Figure 6. Schematic diagram of MDL data fusion

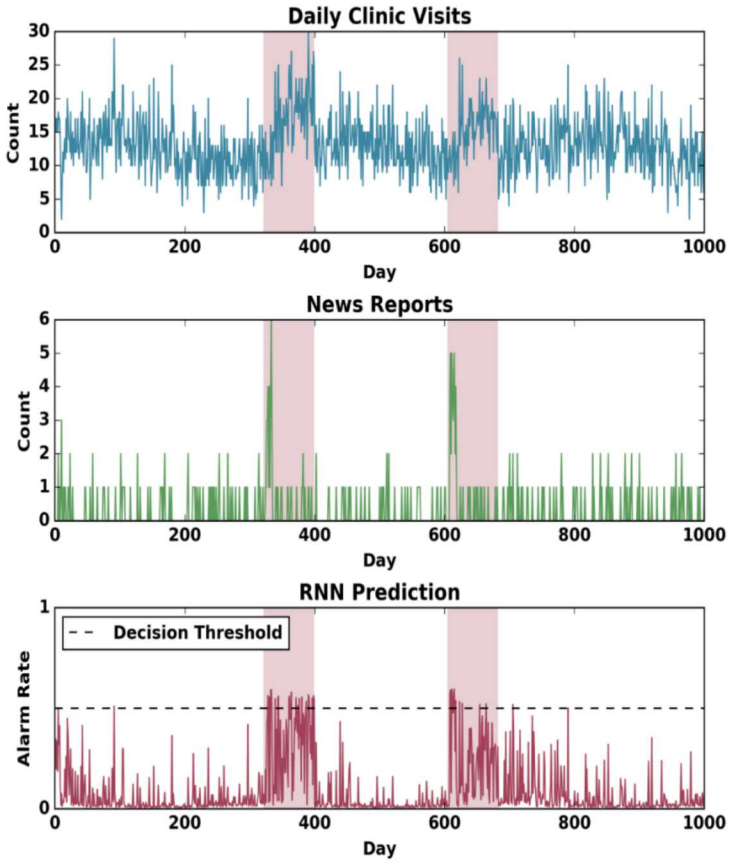

Figure 5. MDL data fusion results. Top two panels show synthetic time-series input data. Bottom panel shows output from example MDL data fusion system

4.5.2. Mixed-Mode Datasets

A recent study demonstrates that multimodal data fusion using recurrent neural networks (RNNs) can improve prediction of patient outcomes from hospital emergency department admission data. The study examamined the predictive accuracy of a series of text classification methods on a deidentified biosurveillance dataset. Next, highest performing text classification method incorporated into a multimodal data fusion configuration, and the quality of the data-fusion derived prediction was compared to previously determined metrics.

4.5.2.1. Biosurveillance Data Set

This study relied upon a deidentified dataset of approximately 1,130,000 emergency department visits in New Mexico during 2016. The dataset was provided by the New Mexico Department of Health Epidemiology and Response Division. Each entry in the data set consisted the chief complaint that each patient reported to upon admission to the ED. Chief complaints are typically short text passages describing what the patient told the admitting ED clerk or nurse. The chief complaints are typically 8 to 20 characters, often with medical jargon and abundant abbrievations and misspellings.

Some typical chief compaints from the New Mexico data set include:

- 16 YR WCC
- EAR
- HEROINE WITHDRAWL

In addition to chief complaints each entry in the New Mexico data set had one or more ICD-10 diagnosis codes assigned to the patient when they were either dismissed from the ED or admitted to the hospital. Additionally, the data set provided coded values to identify the hospital at which the patient presented.

4.5.2.2. Single Mode Text Classification

The dataset was randomly separated into a training set (80% of the data) and a test set (20% of the data). The investigation consisted of determining the accuracy of six natural language processing (NLP) algorithms in predicting a patients discharge diagnosis ICD-10 code based on the chief complaint that the patient provided upon admission to the ED.

Results of the text classification investigation are summarized in Table 1. The table shows classifier accuracy for eleven different text classification methods: term frequency (TF), term frequency-inverse document frequency (TF-IDF), single value decomposition/latent semantic analysis (SVD/LSA), random forest, recurrent RNN with no text embedding (Free), random word embedding, word2vec embedding, GloVe embedding, text2vec/long-short term memory (T2V-LSTM), tweet2vec gated recurrent unit (T2V-GRU), and character to word (C2W). The best performing classification method of the eleven tested is random text embedding/RNN with an accuracy 0f 0.837 as indicated by the bold font on Table 1.

Predictive performance of a range of text classification approaches was investigated by Scott et al. (2018) on a much larger chief complaint dataset from New York. Reported results were consistent with those from the New Mexico data set (Table 1 and Figures 3 – 4).

4.5.2.3. Multimodal Data Fusion

The "Fusion" rows at the bottom show the improved prediction accuracy possible when facility ID is added to the neural network classifier. Adding the additional input increases the prediction accuracy from 0.837 to 0.849. Thus, multimodal data fusion enables greater prediction accuracy than possible from simple text classification of the chief complaint alone.

Table 1. Accuracy of predicted discharge codes derived from text classification

		Accuracy	ROC AUC	Avg. RPC
BoW	Term Freq.	.780	.858	.919
	TF-IDF	.751	.856	.921
	SVD / LSA	.762	.857	.921
	Random Forest	.782	.879	.934
Word	Free	.817	—	—
	Random	**.837**	**.900**	**.944**
	word2vec	.831	—	—
	GloVe	.833	—	—
Char.	T2V LSTM	.820	—	—
	T2V GRU	.820	—	—
	Trained C2W	.816	—	—
Fusion	Fixed	**.849**	**.912**	**.951**
	Free	.849	—	—

The performance of the different text classification and data fusion methods on the New Mexico data set are shown graphically in Figure 7. Using two other metrics of prediction quality rather than accuracy, it is apparent that RNNs (orange line) are the best performing single mode classifier, whereas multi modal data fusion approaches (gray line) are consistently higher performing than any single-mode method.

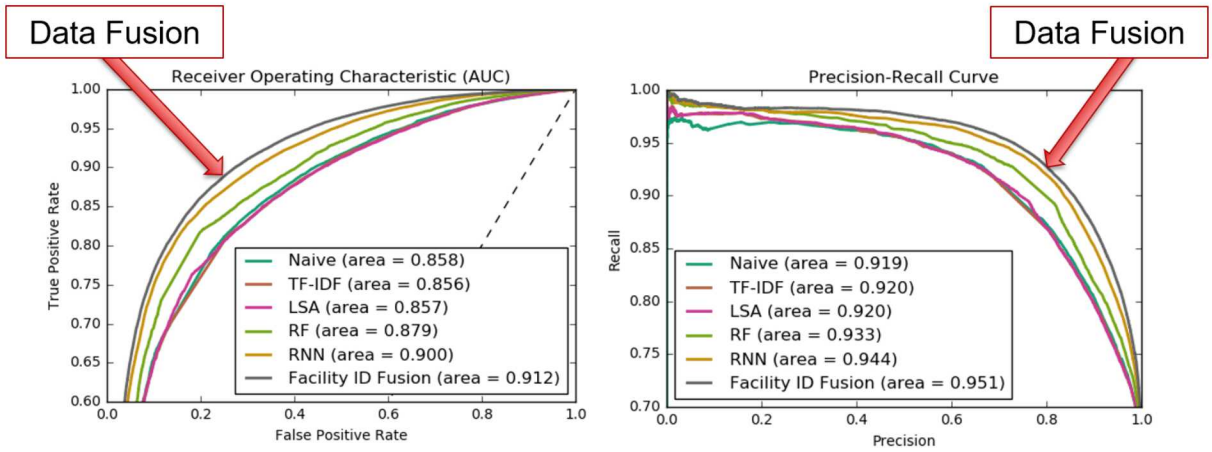

Figure 7. Chief complaint classification performance measured by ROC AUC and PRC.

4.6. Future Research Directions

While this prototype implementation of a functioning MDL data fusion systemis straightforward, implementing a pilot-scale application will require careful design and testing of alternative architectures to determine the most appropriate system. The eventual implementation design and testing process should investigate the relative performance impacts of different ways of structuring the neural networks which underlie the system. Among the elements to be assessed during system development are the timing of the data fusion operation and tuning of architectures for specific data types.

4.6.1. Early vs Late Fusion

A fundamental design decision in developing a pilot-scale MDL data fusion system for a given domain is whether to combine the data feeds before the neural network modeling or after (Figure 8). As discussed in detail earlier, early fusion entails concatenating multiple data feeds into a single composite representation which is then input to a single neural network for processing. Late fusion, on the other hand, typically inputs data from each individual feed into a separate neural network component for learning and classification. The resulting outputs of these individual source-specific neural networks are then combined to generate a consensus indicator of whether or not the combined signals indicate an outbreak. Recently, researchers have proposed hybrid fusion systems that incorporate the advantages of both early and late fusion in the final anomaly detection or decision classification.

Given the potential impact of the selection of early, late or hybrid fusion architecture, further system design should incorporate test implementations of each type of fusion system, followed by rigorous testing on data sets to determine the configuration yielding the best performance for a given input. These initial exploratory design evaluations will help ensure that the fusion architecture selected for inclusion in the final system is well-vetted for the biosurveillance data environment.

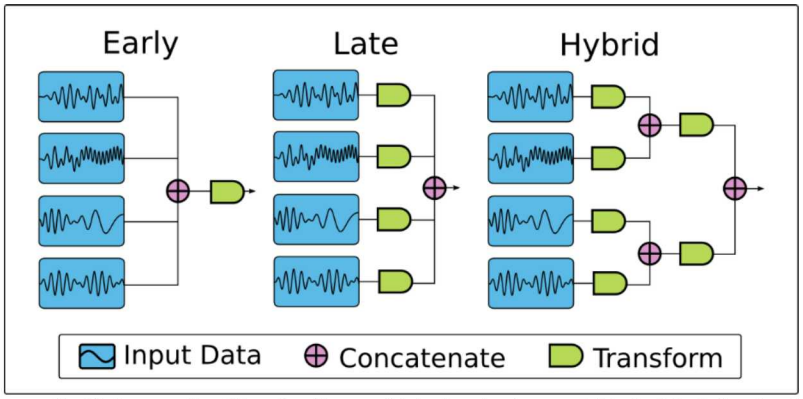

Figure 8. Schematic Illustration of Early, Late and Hybrid data fusion.

4.6.2. Tuning for Specific Data Types

A planned pilot-scale data fusion application will use LSTM-RNNs to handle the data fusion and anomaly detection actions. These networks are configurable in terms of the number of neural units within each network and how they are interconnected. In addition to these network-specific configuration decisions, the LSTM-RNNs are characterized by a number of hyper-parameters, settings which are used to tune the networks for a given function. It is expected that LSTM-RNNs

for different input data types will need to be configured and tuned separately to ensure optimal performance of the network on that data set. As part of the initial design work of this effort, systematic exploration of the effects of alternative network configurations and hyper-parameter settings will be generated for each specific type of input data to be incorporated into the system, e.g. integers, floating point numbers, categorical variables, and unstructured text.

5. DISCUSSION

Automated electronic biosurveillance holds great promise to improve both public health response to natural disease outbreaks, and minimizing potential casualties from a bioweapons attack. In both cases, however, effective monitoring of data streams to reliability generate timely and accurate response to natural disease outbreaks and potential bioweapon attacks face a large number of technical, administrative and policy challenges. In addition to algorithmic anomaly detection and data fusion challenges addressed in this report, collecting raw data on a national scale, aggregating data from across large geographic and political areas, and archiving and storing vast number of records needed is a monumental task. As an outgrowth of the Anthrax attacks in the United States in 2001, the US government has invested much time and money into designing and constructing a comprehensive national system to monitor hospital emergency department (ED) data, first under the guise of BioSense, and subsequently as part of the NSSP program. As of November, 2018 the national biosurveillance program in the United States collects a range of information ED visits from over 80% of the nation's hospitals on a near-real-time basis and, rapidly compiles curates the data in electronic formats amenable to algorithmic processing needed to quickly and accurately detect indicators of anomalous ED activity indicative of disease outbreak from natural or adversarial casuse. However, the size of the assembled near-real-time biosurveillance dataset create unique challenges for anomaly detection routines, as do the inherent noise in the multichannel data.

The study documented in this report has explored a family of data analytics techniques with potential to improve the performance of the existing national biosurveillance system in the United States. Based on the assumption that biological and physiological processes must contend with similarly large and noisy information streams, the analytical approaches that were investigated in this study borrow heavily from natural systems in their design and implementation. The first family of technicques explored in study are modeled after processes that the human immune system uses to detect potential dangerous pathogens and marshall defenses to fight of the potential infection. Similarly, deep learning neural networks adopt motifs from the human nervous system to enable uniquely powerful methods to identify anomalous values in single and multiple data channels that may indicate disease outbreaks. Applying these bio-inspired methods to national biosurveillance represents an innovative approach to define potentially superior analytics methods that could potentially generate much more actionable information on nascient outbreaks, and thus improved return on investment on the national biosurveillance infrastructure.

5.1. Immune system analogs

Immune-system based anomaly detectors and data fusion systems described in section 3 possess characteristics that make them particularly intriguing for use on production biosurveillance systems. The immune-system anomaly detectors are by design specifically tuned to detect novel pathogens. Mimicking processes within the thymus gland where T-cells are generated, the synthetic T-cell anomaly detectors described by Levin and Finley (2017) learn to ignore signals they have seen many times before and to specifically fire when an anomalous signal is detected. Thus the inclusion of T-cell-analog based anomaly detectors in national scale biosurveillance systems could serve to provide an analytical method capabile of detecting novel pathogens, either naturally emerging or specifically engineered as weapons. Since truly robust national biosurveillance would require the capability to trigger on novel pathogen signatures as well as warning of expected seasonal outbreaks that are more virulent than anticipated, an ensemble of immune-system based detectors combined with traditional statistical methods could potentially improve the coverage of the national biosurveillance system to cover both prototypical use cases.

Additionally, immune-system based anomaly detectors are natively multi-channel. As demonstrated by Levin and Finley (2018a), the T-cell analog anomaly detector can operate on a large number of concurrent data feeds, each of which can be a different data type, e.g. numeric, integer, categorical or indicator. This feature is not available with traditional statistical anomaly detectors. For example, the most-used biosurveillance anomaly detection algorithms CUSUM and EMWA are strictly applicable to single channel data streams. Thus the T-cell analog anomaly detectors could be expected to adaptively operate on as many data channels are available for a given region, exploiting the enhanced accuracy potentially available when additional channels are available, but still producing consistent detection performance in areas where datasets are more sparse.

As shown by Flanagan et al. (2018) immune system concepts can also provide useful analyses beyond anomaly detection that are applicable to national biosurveillance. Following the biological analogy that naturally occurring T-cells have properties useful as biosurveillance anomaly detectors, Flanagan postulated that examining mechanisms used by the immune system at a larger scale could likewise inform biosurveillance. In particular, the immune system is functionally organized with distributed hubs of biochemical processing localized in throughout the body in a network of lymph nodes. From an abstract view, lymph nodes serve as hubs of information processing, caching antibodies and directing them to respond to signals received from T-cells in remote tissues. The authors argued that an adaptive national biosurveillance systems could best be thought of as (1) mimicking T-cells in the function of biosurveillance anomaly detection algorithms, and (2) mimicking the network of lymph nodes when considering how data collection, anomaly detection, and response are organized into municipal, county and state public health departments, receiving information specific to their spatio-temporal portion of the larger system and using this local knowledge to instantiate effective control actions. This work cited recent complex system approaches to better understand immune systems and ant colony foraging behavior could be leveraged to derive useful measures of the relative effectiveness of distributed information processing relative to more centralized processing in biosurveillance

Extending this analogy, Beyeler and Finley (2018) demonstrated a mathematical modeling approach to explore the trade-off-space between a highly distributed biosurveillance system, with disease detection and response analysis and response concentrated at local public health offices vs a centralized biosurveillance system where all available public health is funneled to a central data collation, analysis and response facility. Initial model runs showed the relative trade-offs between speed of detection, system-wide sensitivity and specificity, and quality of information provided to epidemiologists responding to a detected event. Early results show that each prototypical biosurveillance network topology has particular advantages. Locally focused networks excelled in rejection of false positive indications, since local cumulative historical knowledge could be applied to better understand that some initial anomaly indicative of a small outbreak may not be likely to spread explosively, and thus best be controlled with standard response measures. Models of the centralized systems were found to leverage large geographic scale to alert local public health nodes of distal outbreaks elsewhere in the network, enabling the local public health nodes to implement heightened awareness and rapid response at first sign of spread. The best performing configuraton was modeled as a compound network which incorporated significant local resource deployment along with hierarchical central information distribution activities. Further model calibration activities are ongoing to enable early parameter study results to be validated against large scale biosurveillance data.

6. CONCLUSIONS

This report has documented novel approaches to improving national-scale electronic biosurveillance systems. Noting specific shortcomings in current biosurveillance system performance such as poor response to noisy datasets and non-stationary baselines, two general families of candidate approaches were identified, implemented, and tested against currently used methods. The evidence and interpretations presented I this report suggest strongly that both general approaches identified should be pursued to determine their feasibility for production biosurveillance applications.

6.1. Immune system analogs

The cited publications on immune system analogs for biosurveillance (Levin and Finley 2016, 2017a, 2017b; Levin et al. 2018; Flanagan et al. 2018; Beyeler and Finley 2018) paint a compelling case for additional research on the feasibility of applying such methods to national scale biosurveillance.

T-cell mimicking anomaly detectors respond reliably to spiked synthetic biosurveillance datasets, often outperforming standard statistical detector performance, particularly in difficult-to-resolve instances of nonstationary baselines and high noise content (Levin and Finley 2016, 2017a, 2017b). Additionally this family of detectors is natively multichannel and multimodal, thus serving not only as a potentially superior approach to single-channel anomaly detection, but also as a superior data-fusion method. This is ability to simultaneously address both anomaly detection and data fusion deficiencies is particularly compelling, given that a recent review on pressing research needs in biosurveillance noted that improving data fusion capabilities of production biosurveillance systems remains a pressing need and recommended as a top research priority (Hopkins et al. 2017).

Pioneering system studies by Flanagan et al. (2018) and Beyeler and Finley (2018) demonstrate the potential utility of considering the design and operation of a large scale adaptive biosurveillance to be a complex adaptive system, and thus amenable to a formidable range of analytical and simulation tools. Flanigan et al. (2018) showed the potential applicability to biosurveillance of prior studies on ant foraging behavior and lymph node networks as exemplars of distributed information and decision making frameworks which operate according to well-understood priniciples.

The mathematical modeling framework developed by Beyeler and Finley (2018) shows promise in providing quantitative guidance to design and operation of large scale biosurveillance systems. By modeling the effects of distributed vs centralized information flow and decision making on the ability of biosurveillance systems to detect evidence of potential disease activity by balanced application of the subject matter knowledge of public health practitioners in individual local jurisdictions along with the efficiency and broad global awareness that centralized resources provide.

6.2. Deep Neural Networks

This study applied the rapidly developing techniques of deep neural networks to several pressing problems in biosurveillance (Levin and Finley 2017c, 2018c, Lee et al. 2017 2018). In addition to single channel anomaly detection, multimodal data fusion and text classification were investigated. In each case, deep neural networks, specifically long-short term memory (LSTM) and gated recurrence unit (GRU) recurrent neural networks (RNNs) performed at least as well as standard methods, and often far exceeding presently used statistical methods. Furthermore, this study demonstrated that newer, more sophisticated neural network architectures such as time-aware recurrent, attention, and wavenet variants of RNNs possess features that may enable even better performance once rigorously tested on biosurveillance data sets.

6.3. Future Directions

This study has shown that bio-inspired analytical architectures can outperform traditional statistical methods on objective comparison tests on synthetic and actual biosurveillance datasets. Further research and development efforts are needed to provide necessary calibration, verification, and validation studies to confirm that the performance improvements described in this study are consistent across the range of conditions in which production biosurveillance systems would be anticipated to function.

Specific recommendations for future work include:

1) Comprehensive testing and evaluation of the algorithms discussed in this report on a wide range of synthetic and production biosurveillance data sets to further verif and calibrate (access to sufficient biosurveillance datasets during the course of this study precluded incorporating this step into the current research)
2) Pilot test algorithms on a high-volume dataset equivalent to the national biosurveillance feed to test algorithms for scalability
3) Identify opportunities to perform anomaly detection and data fusion using the studied algorithms alongside the existing national surveillance to compare sensitivity, specificity and timeliness of the novel methods
4) Apply these novel algorithms in domains other than biosurveillance, such as in deriving longitudinal risk predictions for various ailments from electronic health records from large medical systems

APPENDIX A. APPENDIX PUBLICATIONS AND REPORTS

Following is an abbreviated list of some of the publications and reports from which this report was compiled. These entries include works on immune system and and neural network methods for use in biosurveillance:

Beyeler W, Finley P. Modeling distributed information processing and decision making in national-scale biosurveillance systems. Sandia National Lab.(SNL-NM) (No. SAND20XX), Albuquerque, NM (United States); 2018. (in review)

Finley P, Levin D, Tutorial on Natural Language Processing, Word Embeddings, and Deep Learning for Health Surveillance International Disease Surveillance Annual Conference, Atlanta GA, December 5-8, 2016. (Invited)

Finley P, Levin D. Future directions in NLP for Biosurveillance: Text Embedding and Deep Learning. Presented at 2016 ISDS Annual Meeting, Atlanta GA, December 5-8 2016.

Flanagan T, Beyeler W, Levin D, Finley P, Moses M, Movement and spatial specificity support scaling in ant colonies and immune systems: Application to national biosurveillance., in Evolution, Development and Complexity - Multiscale Evolutionary Models of Complex Adaptive Systems, ed Georgiev G, Smart J, Price M, Martinez C, 2008, Springer (in press)

Hopkins RS, Tong CC, Burkom HS, Akkina JE, Berezowski D, Shigematsu M, Finley PD, Painter I, Gamache R, Del Rio Vilas VJ, Streichert LC, A Practitioner-Driven Research Agenda for Syndromic Surveillance, Public Heath Reports, (in review).

Lee S, Levin D, Thomas J, Finley P, Heilig C. Exploring the Value of Learned Representations for Automated Syndromic Definitions. Online Journal of Public Health Informatics. 2018 May 17;10(1)

Lee SH, Levin D, Finley P, Heilig CM. Chief complaint classification with recurrent neural networks. Journal of Biomedical Informatics 2018 (in Press).

Levin D, Finley P, Evaluating Text Embedding Schemes for Medical Chief Complaint Classification, Sandia National Laboratories 1st Annual MLDL Conference, August 9st, 2017

Levin D, Finley P, Time Aware RNNs for Medical Risk Prediction, Sandia National Laboratories 2nd Annual MLDL Conference, August 1st, 2018

Levin D, Finley P, Time Aware RNNs for Suicide Risk Prediction, Workshop on Outcomes Research in Veteran Suicide, Lawrence Berkeley National Labs, July 30th 2018 (Invited)

Levin D, Finley P. A Spatial Biosurveillance Synthetic Data Generator in R. International Disease Surveillance Annual Conference, Atlanta GA, December 5-8, 2016. Online journal of public health informatics. 2017;9(1).

Levin D, Finley P. A Spatial Biosurveillance Synthetic Data Generator in R. Presented at 2016 ISDS Annual Meeting, Atlanta GA, December 5-8 2016.

Levin D, Finley P. Synthetic data generators for the evaluation of biosurveillance outbreak detection algorithms. Sandia National Lab.(SNL-NM) (No. SAND2018-11533), Albuquerque, NM (United States); 2018.

Levin D, Moses M, Flanagan T, Forrest S, Finley P. Negative selection based anomaly detector for multimodal health data. In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on 2017 Nov 27 (pp. 1-7). IEEE. (Invited)

REFERENCES

Akita R, Yoshihara A, Matsubara T, Uehara K. Deep learning for stock prediction using numerical and textual information. Computer and Information Science (ICIS), 2016 IEE/ACIS 15th International Conference on 2016 Jun 26 (pp. 1-6). IEEE.

al Dosari, M. Unsupervised Anomaly Detection in Sequences Using Long Short Term Memory Recurrent Neural Networks. (Master's thesis, George Mason University).

Ayara M, Timmis J, De Lemos R, Forrest S. Immunising automated teller machines. InInternational Conference on Artificial Immune Systems 2005 Aug 14 (pp. 404-417). Springer, Berlin, Heidelberg.

Beyeler W, Finley P. Modeling distributed information processing and decision making in national-scale biosurveillance systems. Sandia National Lab.(SNL-NM) (No. SAND20XX), Albuquerque, NM (United States); 2018. (in review)

Brownlee J. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. Deep Learning, July 2016

Burkom H, Ramac-Thomas L, Babin S, Holtry R, Mnatsakanyan Z, Yund C. An integrated approach for fusion of environmental and human health data for disease surveillance. Stat Med. 2011 Feb 28;30(5):470–9.

Burkom H. Biosurveillance applying scan statistics with multiple, disparate data sources. Journal of Urban Health. 2003 Mar 1;80(1):i57-65.

Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. arXiv preprint arXiv:1606.01865. 2016 Jun 6.

Cheng KE, Crary DJ, Ray J, Safta C. Structural models used in real-time biosurveillance outbreak detection and outbreak curve isolation from noisy background morbidity levels. Journal of the American Medical Informatics Association. 2012 Oct 4;20(3):435-40.

Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association. 2016 Aug 13:ocw112.

Corberán-Vallet A. Prospective surveillance of multivariate spatial disease data. Statistical Methods in Medical Research. 2012 Oct;21(5):457–77.

Finley P, Levin D, Tutorial on Natural Language Processing, Word Embeddings, and Deep Learning for Health Surveillance International Disease Surveillance Annual Conference, Atlanta GA, December 5-8, 2016. (Invited)

Finley P, Levin D. Future directions in NLP for Biosurveillance: Text Embedding and Deep Learning. Presented at 2016 ISDS Annual Meeting, Atlanta GA, December 5-8 2016.

Flanagan T, Beyeler W, Levin D, Finley P, Moses M, Movement and spatial specificity support scaling in ant colonies and immune systems: Application to national biosurveillance., in Evolution, Development and Complexity - Multiscale Evolutionary Models of Complex Adaptive Systems, ed Georgiev G, Smart J, Price M, Martinez C, 2008, Springer (in press)

Fricker RD, Banschbach D. Optimizing biosurveillance systems that use threshold-based event detection methods. Information Fusion. 2012 Apr 30;13(2):117-28.

Gajewski KN, Peterson AE, Chitale RA, Pavlin JA, Russell KL, Chretien JP. A review of evaluations of electronic event-based biosurveillance systems. PloS one. 2014 Oct 20;9(10):e111222.

Gandhi A, Sharma A, Biswas A, Deshmukh O. GeThR-Net: A Generalized Temporally Hybrid Recurrent Neural Network for Multimodal Information Fusion. In European Conference on Computer Vision 2016 Oct 8 (pp. 883-899). Springer International Publishing.

George D, Huerta EA. Deep Neural Networks to Enable Real-time Multimessenger Astrophysics. arXiv preprint arXiv:1701.00008. 2017 Jan 4.

Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceedings of the National Academy of Sciences. 2002 Apr 16;99(8):5237-40.

Greensmith J, Aickelin U, Tedesco G. Information fusion for anomaly detection with the dendritic cell algorithm. Information Fusion. 2010 Jan 1;11(1):21-34.

Greensmith J, Twycross J, Aickelin U. Dendritic cells for anomaly detection. InEvolutionary Computation, 2006. CEC 2006. IEEE Congress on 2006 Jul 16 (pp. 664-671). IEEE.

Han SW, Tsui KL, Ariyajunya B, Kim SB. A comparison of CUSUM, EWMA, and temporal scan statistics for detection of increases in Poisson rates. Quality and Reliability Engineering

Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

Höhle M, An R package for the monitoring of infectious diseases. Computational Statistics. 1;22(4):571-82. 2007.

Hopkins RS, Tong CC, Burkom HS, Akkina JE, Berezowski D, Shigematsu M, Finley PD, Painter I, Gamache R, Del Rio Vilas VJ, Streichert LC, A Practitioner-Driven Research Agenda for Syndromic Surveillance, Public Heath Reports, (in review).

Huang R, Tawfik H, Nagar AK. On the use of innate and adaptive parts of artificial immune systems for online fraud detection. InBio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on 2010 Sep 23 (pp. 1669-1676). IEEE.

Kulldorff M, Mostashari F, Duczmal L, Yih WK, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. Statistics in Medicine, Volume 26, Issue 8. 15 April 2007 1824–1833

Lau EHY, Cheng CKY, Ip DKM, Cowling BJ. Situational Awareness of Influenza Activity Based on Multiple Streams of Surveillance Data Using Multivariate Dynamic Linear Model. PLOS ONE. 2012 May 31;7(5):e38346.

Lee S, Levin D, Thomas J, Finley P, Heilig C. Exploring the Value of Learned Representations for Automated Syndromic Definitions. Online Journal of Public Health Informatics. 2018 May 17;10(1)

Lee SH, Levin D, Finley P, Heilig CM. Chief complaint classification with recurrent neural networks. Journal of Biomedical Informatics 2018 (in Press).

Levin D, Finley P, Evaluating Text Embedding Schemes for Medical Chief Complaint Classification, Sandia National Laboratories 1st Annual MLDL Conference, August 9st, 2017

Levin D, Finley P, Time Aware RNNs for Medical Risk Prediction, Sandia National Laboratories 2nd Annual MLDL Conference, August 1st, 2018

Levin D, Finley P, Time Aware RNNs for Suicide Risk Prediction, Workshop on Outcomes Research in Veteran Suicide, Lawrence Berkeley National Labs, July 30th 2018 (Invited)

Levin D, Finley P. A Spatial Biosurveillance Synthetic Data Generator in R. International Disease Surveillance Annual Conference, Atlanta GA, December 5-8, 2016. Online journal of public health informatics. 2017;9(1).

Levin D, Finley P. A Spatial Biosurveillance Synthetic Data Generator in R. Presented at 2016 ISDS Annual Meeting, Atlanta GA, December 5-8 2016.

Levin D, Finley P. Synthetic data generators for the evaluation of biosurveillance outbreak detection algorithms. Sandia National Lab.(SNL-NM) (No. SAND2018-11533), Albuquerque, NM (United States); 2018.

Levin D, Moses M, Flanagan T, Forrest S, Finley P. Negative selection based anomaly detector for multimodal health data. In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on 2017 Nov 27 (pp. 1-7). IEEE. (Invited)

Lipton Z, Kale D, Wetzel R. Modeling missing data in clinical time series with rnns. arXiv preprint arXiv:1606.04130. 2016 Jun 13.

Ma L, Lu Z, Shang L, Li H. Multimodal convolutional neural networks for matching image and sentence. In Proceedings of the IEEE International Conference on Computer Vision 2015 (pp. 2623-2631)

Malhotra P, Vig L, Shroff G, Agarwal P, Long short term memory networks for anomaly detection in time series. In Proceedings. Presses Universitaires de Louvain, p. 89, 2015.

Mnatsakanyan Z, et al., "Bayesian information fusion networks for biosurveillance applications." Journal of the American Medical Informatics Association 16.6: 855-863. 2009

Mnatsakanyan Z, et al., "Distributed information fusion models for regional public health surveillance." Information Fusion 13.2: 129-136. 2012.

Morton AP, Whitby M, McLaws ML, Dobson A, McElwain S, Looke D, Stackelroth J, Sartor A. The application of statistical process control charts to the detection and monitoring of hospital-acquired infections. Journal of quality in clinical practice. 2001 Dec;21(4):112-7.

Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11) 2011 (pp. 689-696).

Nguyen D, Joty S, Imran M, Sajjad H, Mitra P. Applications of Online Deep Learning for Crisis Response. arXiv preprint arXiv:1610.01030. 2016 Oct 4.

Nossal GJ. Negative selection of lymphocytes. cell. 1994 Jan 28;76(2):229-39.

Paul M, Held L, Toschke A. Multivariate modelling of infectious disease surveillance data. Statistics in Medicine. 2008 Dec 20;27(29):6250–67.

Ray J, Brownstein J. Nowcasting influenza outbreaks using open-source media reports. Sandia National Laboratories report. Jan 2013

Schiöler L, Frisén M. Multivariate outbreak detection. Journal of Applied Statistics. 2012 Feb 1;39(2):223–42.

Shmueli G, Burkom H. Statistical challenges facing early outbreak detection in biosurveillance. Technometrics. 2010 Feb 1;52(1):39-51.

Shmueli G, Fienberg SE. Current and potential statistical methods for monitoring multiple data streams for biosurveillance. InStatistical Methods in Counterterrorism 2006 (pp. 109-140). Springer, New York, NY.

Socher R, Karpathy A, Le Q, Manning C, Ng A. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics. 2014 Apr 30;2:207-18.

Unkel S, Farrington C, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2012 Jan 1;175(1):49-82.

Vera do Carmo C, Lopes L, Souza A. Comparative study of the performance of the CuSum and EWMA control charts. Computers & Industrial Engineering. 2004 Jul 1;46(4):707-24.

Vohra R, Goel K, Sahoo J. Modeling temporal dependencies in data using a DBN-LSTM. Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on 2015 Oct 19 (pp. 1-4). IEEE.

Woodall WH. The use of control charts in health-care and public-health surveillance. Journal of Quality Technology. 2006 Apr 1;38(2):89-104.

Wu Z, Jiang Y, Wang X, Ye H, Xue X, Wang J. Fusing Multi-Stream Deep Networks for Video Classification. arXiv preprint arXiv:1509.06086. 2015 Sep 21.

Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao R. Deep Learning and Its Applications to Machine Health Monitoring: A Survey. arXiv preprint arXiv:1612.07640. 2016 Dec 16.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Technical Library	9536	libref@sandia.gov

This page left blank

This page left blank

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.