Towards an LSTM-based Approach for Detection of Temporally Anomalous Data in Medical Datasets

(Research-in-progress)

Michael Bowie, Edmon Begoli, Byung Park, Jeevith Bopaiah
Oak Ridge National Laboratory
{bowiemb,begolie,parkbh,bopaiahj}@ornl.gov

Abstract: Health-related data is complex, heterogeneous, and frequently temporally discontinuous which makes it difficult to analyze, and even more difficult to assess for quality, detect errors, or address anomalies. Undetected errors and anomalies within medical data, if not properly addressed, can have far reaching humanitarian and/or financial consequences.

Many machine learning methodologies applied towards automatic detection of complex errors or anomalies are well known. However, the additional challenge in detection of anomalies in health care data is the so-called temporal or contextual dependency – i.e., the challenge of distinguishing "episodes" of anomalous data that are anomalous either in the context of a particular time window or specific to some particular medical condition. In this paper, we present research in progress on the application to medical data of state-of-the-art anomaly detection techniques based on Long-Short-Term-Memory (LSTM) neural networks. We hypothesize that LSTMs present a robust and flexible approach to the temporal anomaly detection in medical data sets due in part to their ability to retain information about both long-term and short-term dependencies on the model outcome.

Keywords: anomaly detection, time series, medical data, long-short-term-memory network, LSTM

INTRODUCTION

The introduction of Electronic Health IT systems has enabled vast collection and efficient management of patient records by storing patient-related information in a digital format. This type of a system allows for the maintenance of a ground truth of patients' records in contrast with paper based patient notes and other health records. However, the introduction of electronic health records has introduced issues that were not as present in the paper based system [BHCW10, BFD16]. Typically, these issues are typographical errors, software induced errors, communication errors, etc. These errors, either human or machine generated, manifest themselves as duplicate information being recorded in multiple health care departments, missing values caused by the values not being available at the time of entry, missing values due to system or human errors, duplicate, incorrect, or contextually incorrect values, and failure to record data, as it becomes available post factum.

Errors and Anomalies

Falsely interpreting anomalous data as expected (normal), or expected data as anomalous, can have critical consequences in the medical domain. For that reason, we treat data of suspicious quality as anomalous until shown as normal. To formalize this, we define an anomaly as a data that deviates from what is standard, normal, or expected [OXF]. Our tasks is then to reliably identify these data deviations from normal data. Consequently and beneficially, we can also leverage anomaly detection techniques in order to progress towards detection and classification of anomalies as data quality errors. As a first step, we taxonomize the anomalies that we encounter into the following standard categories:

- **Point anomaly:**
  A point anomaly is a deviation of a particular data instance from the normal pattern and ranges of the dataset. For example, an average, normal blood sugar level is 80 mg/dl when measured in the morning, or up to 180 mg/dl postprandial (after meal) [AMI16]; anomalous measurements such as 15 or 1000 could likely indicate an error in measurement. Point anomalies can also encompass data quality errors such as switched values, randomly inserted values, and missing values since these would still fall under the definition of a data instance that deviates from the normal pattern. These anomalies are the simplest form of anomalies to detect, and most anomaly detection research revolves around detecting point anomalies.

- **Contextual anomaly:**
  Contextual anomaly is an instance of data being considered an anomaly in a particular context, but not in another context; for example, higher blood sugar measured after a meal would not be anomalous due to context [AMI16]. However, high blood sugar in the morning or before a meal may be anomalous due to context. Again, this can also encompass data quality errors if a data instance is randomly spiked, and it is only detected when it occurs out of context.

- **Collective anomaly:**
  Collective anomaly is an anomalous situation when a collection of similar data instances are behaving anomalously with respect to the entire dataset. For example, if the blood sugar is high for a long period of time, this could indicate an underlying phenomenon; however, one high blood sugar measure in itself is not considered anomalous [AMI16]. Again, this can also encompass data quality errors. For example, if a machine begins to produce wrong outputs, this can be defined as a collective anomaly.

  Interestingly, point anomalies and collective anomalies can be transformed into contextual anomalies if there is sufficient context [CBK09]. Thus, in time-series data, we can use time as the underlying context for the rest of the data. Therefore, if we can successfully detect contextual anomalies with respect to time, then we can simultaneously detect point and collective anomalies.

With this taxonomy in mind, we proceed to construct our approach for the treatment and detection of anomalies in medical datasets.

BACKGROUND

The Department of Energy, along with the Oak Ridge National Laboratory, has partnered with the Department of Veteran Affairs (VA) in their Million Veteran Program (MVP) [MVP], a voluntary research initiative that aims to analyze the largest repository of health care data in the world, which includes Electronic Health Records (EHR) for 22.5 million individuals, and genomic data from over 560,000 Veterans [BKB16]. While this is a true "goldmine" for research, the data inherently carries a number of data quality issues and possible errors. One of the problems that have been identified as important and critical for advancement of the state of medical research under this initiative is the need for timely and accurate detection of data quality anomalies and errors occurring within VA's Electronic Health Records.

Manifestations and Impacts of Medical Data Errors

In this work, we choose to systematically understand the data quality errors by first treating them as anomalies. We do so because, in practice, it is not always clear whether a new data instance is an error, or if it is simply a new observation. Hence, applying automated anomaly detection techniques provides an alternative method of detecting unusual variations, and alerting users to what might be an error. Data errors, frequently introduced into the EHRs at the point of data entry, propagate through the different stages of clinical procedures and may sometimes directly or indirectly interfere with the patient's therapeutics. One such instance, that could have resulted in wrong medication if not for the acute observation of the pharmacist is as quoted in [INS15]: "*The patient's weight was entered as 99kg in the EHR system. When pharmacy called the care unit to confirm the weight to dose an antibiotic, the nurse stated that the correct weight was 49kg. The correct weight was used to calculate the dose."* This is just one illustration; there have been numerous such occurrences, which include transposition of height and weight, incorrect age, typographical errors in prescribed drugs and their dosage, ordering for a wrong lab test, etc. Some of the error types observed in the EHR are categorized as: [WSF15]

- Substitution Errors: e.g. substitution timestamps "23:59" instead of 11:59PM.
- Missing Errors: occurs when end users do not document activities or delete data.
- Random Errors: random values can occur when values are manually inserted by the clinician.
- Systematic bias: when a value is shifted due to a system error, e.g. an erroneous computer clock that is undetected.

In some cases, patients were assigned incorrect ICD codes, either on accident, or with the intention of falsely increasing the billing for the procedure or treatments performed for the patient. Such errors can cost the health care systems billions of dollars either in direct or indirect costs including damages. Andel et al., [ADHM12], report that "In 2008, medical errors cost the United States $19.5 billion."

While these errors and anomalies can have significant humanitarian or fiscal impacts, detecting them can be very challenging due to the complexity of detection.

Complexity of Detection

Most anomaly detection techniques only detect point anomalies because it is often too challenging for many of the typical algorithms to detect complex, multi-variate, and temporally heterogeneous kinds of anomalies. One reason for this is that labeled data typically belongs to healthy or normal patients; thus, most techniques must be semi-supervised or unsupervised [CBK09]. Typically, in the process of analysis and in order to detect complex anomalies, data scientists need to apply many different techniques for the different kinds of possible anomalies: point, collective, and contextual. Moreover, contextual anomalies typically require extensive domain expertise to interpret, which can be very time intensive. For instance, the patients' electronic health records (EHRs) consist of temporally dependent data points, meaning that time creates an important context to whether a data point is normal or anomalous. For example, very low sodium readings for an otherwise healthy patient would be an anomalous state. However, if a patient had their large intestine removed in the past, then a low sodium reading for this patient is not as anomalous. In this example, we have the temporal context of a past procedure that impacts today's lab results. A simple clustering technique that ignores time would incorrectly label the second patient's results as anomalous. Thus, since time provides a high level of context, it would be a severe loss of information to ignore this time as an important variable in anomaly detection. Thus, opportunities exist for better anomaly detection techniques in the context of medical data that can detect all three kinds of anomalies mentioned in the preceding section. We propose that a neural network based approach to anomaly detection, specifically Long-Short-Term-Memory networks, can go a long way towards automatically solving these challenges.

A Proposed Role for Long-Short-Term-Memory (LSTM) Networks

Long-Short-Term-Memory networks, or LSTMs, were shown to be effective in automation of some of the challenges associated with temporally dependent anomaly detection problems. With that observation in mind, and with cognizance of the need for LSTMs to be applied correctly, the goal of the research in progress presented in this paper is to leverage LSTMs to detect both data quality errors and the three kinds of anomalies described above. Data quality errors such as substitution, missing values, and switched values can also be detected because, as previously defined, they stand out as anomalies when compared to normal data in large datasets. The unique advantage of LSTMs is that they are capable of 'remembering' the variable inputs even from the very beginning of a sequence, and they can use these inputs to predict sequences much further into the future. Previously, with basic recurrent neural networks (RNNs), while processing the streams of data, the network would begin to "forget" about the initial inputs. This problem is called the *vanishing gradient problem*. LSTMs were developed precisely in order to solve these problems. More technically, LSTMs have a series of 'gates'. These gates have functions that decide how much of the 'memory' or influence of the previous inputs to 'let through' or weigh into the calculation for the output at a particular time-step. Most LSTMs have at least three of these gates called an input gate, a forget gate, and an output gate. Each gate will learn to determine how much information to persist through the cell state, or the main information flow in the network [OLA15].

For example, this means that an LSTM can 'remember' that a particular patient had a procedure to remove their large intestine; thus, it can learn that low sodium values from that point on are not as anomalous as they might appear for average cases. Additionally, when the LSTM predicts a sodium value for a different patient who still has their large intestine, the LSTM will learn to use its forget gate to 'forget' the previous patient's unique variables. This means that the fact that the previous patient had their large intestine removed would not be taken into account into the prediction for a patient who still has their large intestine. We find this helpful type of learning to be much more difficult to automate with other adaptive learning algorithms.

LSTM Architecture

Neural networks are a way of finding solutions, or function mappings, from *x* to *y*. For example, a common question asked of a neural network is *if given x: a picture of a cat, can we produce y: the text-label 'cat'?* Instead of explicitly programming software with thousands of rules or functions to pick out certain characteristics of a cat, we can provide a program with thousands of examples of labeled cat images, so that the program can learn its own rules or functions to label an image successfully.

The most basic neural network is a *feed-forward neural network* where the data moves forward in only one direction to map one input to one output. Therefore, each input and output pairs are independent of all other inputs and outputs. These feed-forward networks work great for classification and regression tasks. However, if we have a forecasting task where we want to predict where the sequence will be in the future, it is helpful to know where the sequence has been in the past.

Recurrent neural networks (RNNs) were invented to solve this problem. Recurrent neural networks have loops that allows information to persist. However, as previously mentioned, basic recurrent neural networks suffer from the *vanishing gradient problem*.

LSTMs were invented to solve this problem by retaining the long-term knowledge about some features of data.

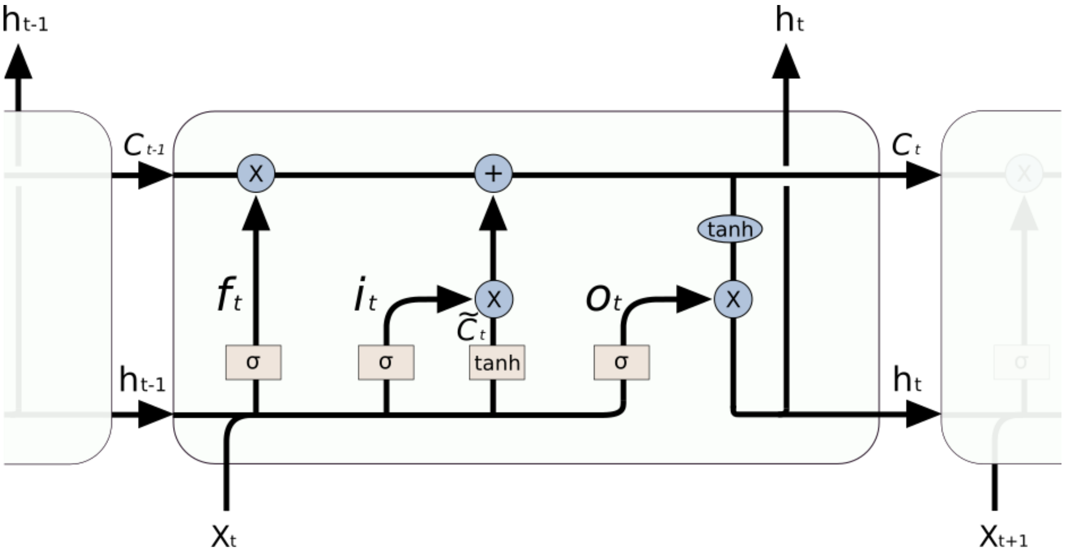

Figure 1: LSTM Architecture (OLA15)

The core of an LSTM is the cell state, C_t , represented by the horizontal line at the top of the diagram. Other than a couple minor interactions, information flows unchanged. However, LSTMs can add or remove information to the cell state by the cell gates. These cell gates have corresponding sigmoid layers, σ, that output a number between zero and one. While a zero does not allow any information to pass through, a one allows all information to pass through.

The first gate in this LSTM layout is the "forget gate", f_t. This gate decides what information is kept or discarded from the cell state, and this decision is made by the sigmoid layer. In a diagnosis prediction problem, the model may want to factor-in gender and age into the prediction. However, the cell state would want to forget the gender and age of the previous patient to make an unbiased prediction for the next patient.

The next gate decides what new information will be stored in the cell state from the input, $x_{t.}$ This two part decision includes an “input gate", $i_{t,}$ that decides which information will be updated and a *tanh* layer that creates a vector of candidate values, $\tilde{C}_t$, that could be added to the state. In our patient diagnosis example, we’ll want to add the gender and age of the new patient to the cell state to replace the old one.

Finally, the LSTM cell decides what will be output, h_t . This is a filtered version of the cell state that decides which parts will be output. The filtering of the cell state is done by the *tanh* function and the "output gate", o_t , decides how much of the filtered state to let through to the next cell state. For example, if the next step in the LSTM is to calculate the risk of a patient having a particular disease, then the output of the current cell may be a vector of potential diseases that the patient risks of being affected by.

There are several variants of the LSTM model; however, for our experiments, we will be using the basic architecture as detailed here [OLA15].

Anomaly Detection with LSTM

There are two main state-of-the-art techniques for anomaly detection with LSTMs. First is a prediction based anomaly detection. The second is an encoder-decoder, or reconstruction based anomaly detection[MVSA15]. The aim of our research is to investigate which, if any, of these two techniques will work best for the types of problems present in our research.

- **Prediction Based:**
  The LSTM learns how to model the normal sequence without any anomalies or errors. We then

train the LSTM to successfully predict the next sequence. This prediction can then be tested on an erroneous sequence. To detect the anomaly, the error between the prediction and the data instance is measured. If the difference is large, with the gap either arbitrarily or analytically assigned, the data can be classified as a potential anomaly.

- **Reconstruction Based:**
  The LSTM learns how to encode a sequence into a smaller representation of the data and decode, or reconstruct, the exact sequence successfully. In order to detect anomalies, the idea is that when the LSTM does poorly at reconstructing a sequence, then this must be a sequence it has not been seen before. If the model has not seen this sequence before, then it can be classified as a potential anomaly.

Our aim is to test how well an LSTM can detect anomalies, including potential data quality errors, in typical medical data using both strategies.

APPROACH FOR LSTM-BASED DETECTION

The following section describes preliminary results on using the LSTM- based model for temporally and contextually sensitive anomaly detection in medical datasets.

We use LSTM models to detect anomalies in a given progression of *m* variables that constitute data for a patient. Formally, this is a multi-variate time series $S = \left\{s^{(1)}, s^{(2)}, \cdots, s^{(n)}\right\}$, where $s^{(t)} \in R^m$ that is observed at time *t*.

For a prediction based approach, we construct an LSTM model that learns to predict a sequence of the next *l* vectors when a vector $s^{(i)}$ is presented to the model, where $1 \le l \le n$. Formally, let us define such *l* sequences of *m* variables given $s^{(i)}$ as $\left\{\tilde{s}_i^{(i+1)}, \tilde{s}_i^{(i+2)}, \cdots, \tilde{s}_i^{(i+l)}\right\}$. Consequently, for a vector $s^{(i)} \in S$, this model generates error prediction vectors $\left\{\tilde{s}_j^{(i)}, \tilde{s}_{j+1}^{(i)}, \cdots, \tilde{s}_{j+l}^{(i)}\right\}$, if $l \le i \le n - l$. This error is computed by calculating the difference between the prediction and the actual value. A set of such prediction vectors is used to assess the likelihood of $s^{(i)}$ being an anomaly [MVSA15].

For the reconstruction based approach, we construct a slightly different LSTM model that learns to reconstruct $s^{(i)}$ when presented $s^{(i)}$ itself. This reconstruction is completed by an encoder-decoder strategy whereby $s^{(i)}$ is encoded into a vector representation and decoded into a reconstruction of $s^{(i)}$. Again, this model generates error prediction vectors $\left\{\tilde{s}_j^{(i)}, \tilde{s}_{j+1}^{(i)}, \cdots, \tilde{s}_{j+l}^{(i)}\right\}$, if $l \le i \le n - l$. This error is computed by calculating the difference between the reconstruction of $s^{(i)}$ and the actual values belonging to $s^{(i)}$ and the set of prediction vectors is used to assess the likelihood of $s^{(i)}$ being an anomaly [MRA16].

Stacked LSTM Prediction Based Model

The following LSTM network architecture is used as described in [MVSA15]: In order for the architecture to learn higher level temporal features, we stack LSTM layers. This means that each layer is fully connected to the layers above through feed forward connections.

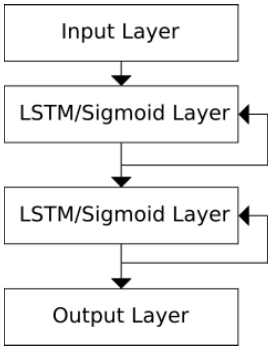

Figure 2: Stacked LSTM Architecture

Data and Experiments

Our current approach to the evaluation of LSTMs involves the use of a realistic, "clean", medical dataset which is inserted with errors of our choosing to assess the detection power and flexibility of the LSTM approach.

Medical Information Mart for Intensive Care Database (MIMIC)

Given the variety of anomalies that could occur in the EHR's, we have identified one such source of system errors in the bedside monitoring devices. In this project, we analyze time series signals such as heart rate, systolic blood pressure, diastolic blood pressure, and respiration rate. Using these signals, we aim to detect the device errors present in them.

We start with a default, standard dataset to train, and an algorithmically altered erroneous dataset to test. The training dataset will simply be the lab results of over 40,000 real patients from the Medical Information Mart for Intensive Care (MIMIC-III) database [JPS16]. These lab result files are CSV files that include all the sequentially ordered lab results the patient received during an ICU stay.

Error Insertion

Since the MIMIC III data that is used in training the neural network model is devoid of anomalies and is considered to be normal data, we represent the anomalous behavior of the EHR by adding random data into the otherwise normal data. Since: (i) the nature of an anomaly in the EHR is not yet well established and (ii) accounting for all anomalous behaviors is a substantial task, we focus on the following strategies for introducing errors into the bedside monitoring signals:

- Random Errors: we add errors at 2%, 4%, and 7% of the dataset. For each of these error instances, the error is calculated by increasing or decreasing the existing value by 50%, 75%, and 100%.
- Mismatch Errors: where we randomly switch values, e.g., diastolic/systolic, at 2%, 4%, and 7% of the dataset.
- Missing Errors: where we delete values at 2%, 4%, and 7% of the dataset.

Based on probability estimates, either continuous errors or point errors are added to the time series. Depending on how our predictive model performs on this set, more complex strategies can be devised as well.

Anomaly Detection Using the Prediction Error Distribution

The following anomaly detection technique is applied as described by [MVSA15]: We compute an error vector that is the difference between the prediction of $s^{(i)}$ at the next time step and the true value at $s^{(i)}$. The error vectors are fit to a multivariate Gaussian distribution $\square = \square(\mu, \Sigma)$. An observation is classified as anomalous if the prediction is less than τ, else the observation is classified as normal. The normal validations tensors are used to learn τ by maximizing the F-score. This method assigns positive scores to anomalous points and negative scores to normal points [MVSA15].

Evaluation

In order to evaluate how well the LSTM performs at correctly detecting these errors, we can use a confusion matrix that compares our true and false positive rate and compares our true and false negative rate. These rates will provide the model's accuracy, precision, and recall metrics. We will test both anomaly detection strategies, prediction based and reconstruction based, with an LSTM, but we will also test common regression and control chart methods. These methods will serve as a baseline to compare the performance of the LSTM-based approach. Note however, implementation of a reconstruction based strategy is still being researched. More detail on this strategy is provided in the next section.

During testing of these models, the most challenging aspect of anomaly detection in the medical domain is that there is a very high cost in classifying an anomaly as normal. As previously mentioned, undetected anomalies and data errors can have a significant negative impact on the patient's health. Thus, we will be optimizing our models for recall, or true labeled positives divided by all positives. This will ensure we detect as many positives as possible. While regression and control charts will be easier to implement and may even detect many anomalies, our prediction is that the LSTM will significantly outperform regression and control charts in detecting the largest percent of anomalies.

Preliminary Results

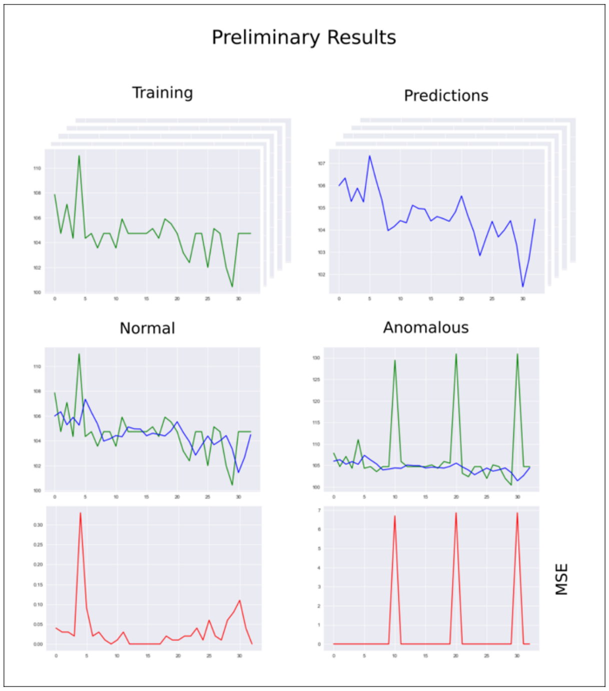

Figure 3: Preliminary results (sample sequences: green, predictions: blue, MSE: red)

The preliminary results are promising as they show the LSTM is learning a predictive model of the sequences. These predictions are compared with the original sequence in order to calculate the mean squared error. This error is currently interpreted as the probability of an anomaly. However, in the future, the mean squared error, or error vectors, will be fit to a Gaussian distribution in order to calculate a likelihood threshold for a potential anomaly, as previously stated. This is important since the normal data may have subsequences that appear anomalous simply due to the nature of the ICU data but in fact are normal.

ONGOING RESEARCH AND EXPECTED OUTCOMES

Parts of this research are still ongoing as there are a few challenges that must be addressed before experimental results can be produced. Anticipated challenges include: a) handling missing values, b) implementing a reconstruction based LSTM, and c) figuring out how to incorporate more data into the models such as demographic and diagnosis information.

Missing Values

Since the MIMIC-III dataset consists of real ICU events of real patients, the data is very sparse. Not all lab results are always taken at each time step, and most patients have a varying length of stay. While some health indicators, such as the heart rate, are monitored constantly, some other, such as the glucose level, are measured for less than 50% of the time steps. To overcome the sparseness in the data, we will initially use imputation strategies that provide a reasonable estimate for the missing data. The records that have more than 50% missing values are not considered in our experiment. The strategy to account for the missing values is to calculate the mean of the existing data, and assume it to be the reasonable prediction for the missing data. In order to overcome the fact that most of the patients have a varying length of stay, we will implement dynamic batch padding to process different time series lengths. Initially, we will test if padding the time series with zeros will be effective in our models.

However, using naive imputation strategies to ’fill in the gaps’ can result in loss of information. As previously mentioned, one of our goals is to determine whether a missing, or randomly deleted value, is anomalous. Obviously if the values are imputed to become the mean of the surrounding values, determining whether a missing value is anomalous becomes impossible. However, missing values are not always predictable. Thus, the prediction based LSTM may not work in predicting missing values. However, according to recent research by [MRA16], reconstruction based LSTM can perform much better in finding unpredictable anomalies in predictable or unpredictable sequences.

Reconstruction Based LSTM Anomaly Detection

Leveraging a reconstruction based LSTM for anomaly detection is still an ongoing part of our research. The benefit of this strategy is that it is robust to predictable and unpredictable series. Thus, this strategy could be leveraged in order to find anomalies in sequences that may not be predictable, or such sequences that include randomly missing values. We predict that we will need to implement a reconstruction based LSTM in order to detect anomalously missing values.

The goal with this strategy is for the LSTM to encode a vector representation of a sequence and learn to successfully reconstruct the sequence from the vector representation. The difference between reconstruction and the original sequence is the error. Thus, if this model is trained on normal data, then the reconstruction error will be low for a normal, or expected sequence. However, the error will be high on sequences the model has never ’seen’ before. These high errors could be classified as potential anomalies [MRA16].

Part of the ongoing research with a reconstruction based LSTM is finding the optimum window size for the model. The window size is the length of the sequences used as input. If an important variable is missing, this could lead to a large error during reconstruction. The goal is to provide a clear ’picture’, or window, to the model in order to provide not only sufficient information for reconstruction, but also serve as a valid representation of a normal state.

Incorporating Additional Data

Another challenge we are trying to address is how to incorporate additional information and features into our models. Potentially, the more data we can use, the better our models will perform and, potentially, the faster our models will train. The features we initially aim to incorporate into the model for analysis are demographic, diagnosis, and procedural information. Incorporating these features can have a huge impact on predictions because lab results can be drastically different based on age, gender, diagnosis, and past procedures. However, the biggest challenge is how to blend these discrete and continuous variables into a single form the LSTMs will accept and understand. Initially, we are testing a basic method of doing this. Since demographic, diagnosis, and procedural data change with time, a basic method of incorporating this would be adding these data to each time series. For example, we would extend the rows of lab values to include a one hot encoding of comorbidities. Thus, the diagnosis information would be tracked by the granularity of the time series. Similarly, we would also extend these rows to include demographic information that would be tracked over time. Hence, at any given time series the row would encompass the state of the patient as a whole. Of course, as the number of patients and time series events grow, this can become memory intensive since we are copying the same information over and over. However, while we research better methods, this can be a basic way of including more information into our anomalous/normal predictions.

FUTURE AND RELATED WORK

Our current research, reported in this paper, inspired us to look further into other potential outcomes that could contribute to the state-of-the-art in the data quality, and anomaly detection for medical data research. We have identified three relevant areas, namely: i) health data quality benchmarking, ii) detection of inconsistencies between structured and textual data, and iii) use of memory augmented networks.

A Need for a Health Data Quality Benchmark

We have seen dramatic progress in machine learning algorithms in a variety of applications. These algorithms are advancing the state-of-the-art in image classification, text mining, speech recognition, and more. Catalysts for these advancements include community recognized benchmarks such as ImageNet Large Scale Visual Recognition Challenge [RDS15]. These benchmarks allow researchers to improve model performance while competing with the community on very specific tasks. This focus fosters collaboration and advances progress. Until recently, the health care domain has lacked community accepted datasets and benchmarks to test domain-specific machine learning algorithms. In 2016, YerevaNN, a non-profit computer science and mathematics research lab, published four benchmarking tasks for the Medical Information Mart for Intensive Care (MIMIC- III) database [HKKG17]. They propose the MIMIC-III database to be a standard benchmarking database for machine learning research for health care. Using the MIMIC-III database for bench- marking solves a big issue in health care related research. Researchers at YerevaNN have published four benchmarks each on different tasks regarding mortality prediction, decompensation prediction, length of stay forecasting, and acute care phenotyping. They even produced a multitask LSTM architecture to solve all four tasks simultaneously. YerevaNN is already leveraging these LSTM architectures with impressive prediction benchmarks on temporally dependent medical data [HKKG17].

Our goal is to contribute to YerevaNN's effort and create the first anomaly detection benchmark for the publicly available MIMIC-III database. Initially, we are focusing our efforts on lab results and detecting data quality anomalies such as impossible values, switched values, and erroneously inserted values. Not only does this focus serve as a proof-of-concept for more complex tests, but it is also immediately valuable for a health care organization. However, we would like to expand this effort to include a wider range of data provided by the MIMIC-III database such as demographic information, comorbidities, and more. By incorporating as much data as possible from the MIMIC-III dataset, we will be able to provide a true anomaly detection benchmark for others to challenge. Perhaps LSTMs will be flexible enough to incorporate this additional data when detecting anomalies and data errors and will serve as a standard for anomaly and error detection in medical data.

Detecting Textual and Structured Data Inconsistencies

Another area for future work is incorporating physician notes into the anomaly detection models. This area of work would aim to address the problem when textual and structured data is inconsistent. For example, there were cases in EHRs where the doctor's notes and the information found in the structured data were either just slightly different or completely contradictory. One source for these errors is the use of copy and paste in EHRs. In order to save time, physicians can copy the notes from a previous visit into the notes of the current visit. This happens especially when the visits are regarding the same diagnosis. However, a problem that occurs is that certain parts of the notes including medication dosage are mistakenly not changed. It has been found in situations where patients almost underwent the same procedure twice simply because the exact same notes were entered for a follow up visit [HOF14]. We posit that natural language processing strategies could be incorporated into a model that compares the free form notes with the structured information in order to determine a match score. If the match score is high, then there is no error. However, if the match score is low then the data would need to be reviewed by a clinician. This kind of automation would significantly reduce these textual and structured data inconsistencies that can arise in EHRs.

A Role for Memory Augmented Networks

Another area for future work consists in adding more learning power and generality to the LSTM or other neural network. We propose a future model structure for anomaly and error detection such as a memory augmented network [GWR16]. Instead of the LSTM only relying on its own internal memory, the LSTM can be given read and write capabilities to an external memory. This external memory would serve as a sort of random access memory to the LSTM, or other model, that serves as a controller. Not only could this allow a model to be tailored specifically to a patient, but this model could also provide a layer of reasoning. This reasoning layer could be used for dynamic cohort creation, for example [RHD16]. Concretely, the external memory could be queried similarly to an SQL query. For example, we could ask the memory to display all patients with rheumatoid arthritis that have anomalies in their EHR. Or, display which clinicians have inserted the most anomalies, or data quality errors, into an EHR.

CONCLUSION

The EHR system allows us to maintain a single source of truth of patients' records. However, the introduction of the electronic format for the health records has also introduced a plethora of data quality issues. Since medical data is complex, voluminous, and it has one of the highest costs for undetected anomalies, reliable automated approaches for detecting these data quality issues are critical. To address these problems, we are exploring state-of-the-art anomaly detection techniques for time series based or time-dependent medical data, specifically LSTM. We will examine which particular LSTM-based anomaly detection technique works best – either the prediction based or the reconstruction based – and we will present our findings in upcoming publications. Our expectation is that LSTM-based approaches will not only outperform common statistical techniques, such as regression and control charts, but will also be the most flexible for addressing new problems.

ACKNOWLEDGMENT

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

The research presented in this paper was supported in part by the joint program MVP CHAMPION between the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development and the Department of Energy.

REFERENCES

[ADHM12] Charles Andel, Stephen L Davidow, Mark Hollander, and David A Moreno. The economics of health care quality and medical errors. *Journal of health care finance*, 39(1):39, 2012.

[AMI16] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278 – 288, 2016.

[BKB16] Begoli, Edmon, Derek Kistler, and Jack Bates. "Towards a heterogeneous, polystore-like data architecture for the US Department of Veteran Affairs (VA) enterprise analytics." Big Data (Big Data), *2016 IEEE International Conference on. IEEE*, 2016.

[BDF16] Begoli, Edmon, Ted Dunning, and Charlie Frasure. "Real-Time Discovery Services over Large, Heterogeneous and Complex Healthcare Datasets Using Schema-Less, Column-Oriented Methods." Big Data Computing Service and Applications (BigDataService), *2016 IEEE Second International Conference on. IEEE*, 2016.

[BHCW10] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM Comput. Surv., 41(3):15:1–15:58, 2009.

[OLA15] Christopher Olah. Understanding lstm networks. 2015.

[GWR16] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626):471–476, 2016.

[HKKG17] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:1703.07771, 2017.

[HOF14] Sharona Hoffman. Medical big data and big data quality problems. 2014.

[INS15] ECRI Institute. Wrong-record, wrong-data errors with health it systems. 2015.

[JPS16] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. Scientific data, 3, 2016.

[MRA16] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148, 2016.

[MVP] MVP - for researchers and research partners. https://www.energy.gov/articles/doe-and-va-team-improve-healthcare-veterans.

[MVSA15] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In Proceedings, page 89. Presses universitaires de Louvain, 2015.

[OXF] Oxford Dictionary: https://en.oxforddictionaries.com/definition/anomalous.

[RDS15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.

[RHD16] Jack Rae, Jonathan J Hunt, Ivo Danihelka, Timothy Harley, Andrew W Senior, Gregory Wayne, Alex Graves, and Tim Lillicrap. Scaling memory-augmented neural networks with sparse reads and writes. In Advances in Neural Information Processing Systems, pages 3621–3629, 2016.

[WSF15] Michael J Ward, Wesley H Self, and Craig M Froehle. Effects of common data errors in electronic health records on emergency department operational performance metrics: A monte carlo simulation. Academic Emergency Medicine, 22(9):1085–1092, 2015.