2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Predicting Hospital Length of Stay using Neural Networks on MIMIC III Data

Thanos Gentimis
Florida Polytechnic University
Lakeland, Florida 33805–8531
Email: agentimis@floridapoly.edu

Ala' J. Alnaser
Florida Polytechnic University
Lakeland, Florida 33805–8531
Email: aalnaser@floridapoly.edu

Alex Durante
Florida Polytechnic University
Lakeland, Florida 33805–8531
Email: alexdurante0262@floridapoly.edu

Kyle Cook
Florida Polytechnic University
Lakeland, Florida 33805–8531
Email: kcook@floridapoly.edu

Robert Steele
Florida Polytechnic University
Lakeland, Florida 33805–8531
Email: rsteele@floridapoly.edu

***Abstract*—In this paper we explore the use of neural networks for predicting the total length of stay for patients with various diagnoses based on selected general characteristics. A neural network is trained to predict whether patient stay will be long (> 5 days), or short (≤ 5 days) as of the time the patient leaves the ICU unit. Our dataset is drawn from the MIMIC III database and all code was written in R and in $Postgress$, while the computations were executed on the Florida Polytechnic University's supercomputer. Our prediction accuracy is approximately 80% and clearly outperforms any linear model.**

I. Introduction

As per the current literature, the use of neural networks [5] to predict the length of stay for any patient, not limiting the network's prediction to just one disease condition has not been attempted in the past.

Predicting hospital length of stay is of substantial value for hospital resource planning and management. Being able to better estimate how much longer a patient will remain in hospital, can assist in scheduling the usage of wards and hospital beds including the scheduling of elective surgeries based on upcoming availability of beds. Length of stay explains 85 to 90 percent of the variation in hospital costs between patients [11].

As such, a model that is able to better estimate length of stay will assist hospital adminstrators, clinicians, patients and payers. For hospitals it is desirable to optimize the use of beds to best provide care, for clinicians predictive models can provide adjunctive clinical decision support, for patients improved planning and prediction can contribute to their quality of care and for payers, who are responsible for paying for healthcare, they are continually seeking tools to increase cost analysis and prediction. In addition the move towards value-based care requires greater prediction and optimization of how to maintain the health of a population. Being able to better predict length of stay and hence better care for patients is an important part of value-based care.

A challenge for such a predictive model is that numerous factors, with not completely understood and complex relationships can affect length of stay. This suggests the potential applicability of a neural network-based approach. The challenge of these multiple factors and their complex inter-relationships has contributed to the lack of a generic predictive system for length of stay to-date.

In addition, historically health records were kept in a non-digitized form. The last decade has seen a significant increase of the use of electronic health records and other digital health information systems. This also contributes anew to support the increased applicability of machine learning approaches and to the increased possibilities for neural network-based health applications.

In our effort we are utilizing the MIMIC III dataset [8] to provide hospital stay data upon which the neural networks are trained. The dataset is freely available and provides detailed data of patients admitted to the Beth Israel Deaconess Medical Center in Boston and was developed by the MIT Lab for Computational Physiology.

After an initial exploration of our dataset using the package "autoweka" from the WEKA machine learning software [4] a neural network was created based on patient characteristics without including specific bio-markers to give an estimate of the stay in the hospital after the patient leaves the ICU. These networks were implemented using the R package neuralnet [13] after the appropriate pre-processing.

The neural network was run 100 times on 100 random subsets of the dataset, with each subset split into a training set and test set. The accuracy results of all runs were recorded.

This article is organized as follows: In section II we describe the general framework of our method giving an overview of the mathematical, statistical and machine learning tools we are using. In section III we describe our experiments, present the results of our methods and give their predictive accuracy. Following that we discuss the results and their significance in

978-1-5386-1956-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191

IV, then describe future work and finally conclude the paper in VI.

A. Literature Review

As stated above, there are no current models that attempt to predict the length of stay of any patient independent of their disease. Many studies, however, have attempted to determine factors that can predict the length of stay where the patients are limited to a certain health condition or group of health conditions. In our review, we have found the following models with varying degrees of data specification and accuracy.

In a paper from Maria Kelly et al., hospital records were analyzed to determine if any factors could predict hospital length of stay (LOS) and readmission after colorectal resection through linear regression [9]. Data was provided by a combination of databases from the National Cancer Registry (NCR) and the Hospital In-Patient Enquiry Scheme (HIPE) that contains records of patients in Ireland. STATA was used to determine the best variables for logistic regression using a combination of likelihood ratio, Hosmer, and Lemeshow tests. For LOS, it was determined that age, higher levels of co-morbidities, and marital status were associated with an increased LOS.

Ying Wang et al. analyzed hospital records to identify predictors of an increased LOS after acute exacerbation of chronic obstructive pulmonary disease (AECOP) [14]. They utilized data received from the Oslo University Hospital, Aker and their analysis involved patients whose main diagnosis or secondary diagnosis was Chronic Obstructive Pulmonary Disease. A Multivariate logistic regression model was created to assess predictors of a LOS that was longer than 11 days (which corresponded to above the 75th percentile of LOS times). Results from the logistic regression show that being admitted between Thursday and Saturday, high $PaCO_2$, low serum albumin level, and having heart failure, diabetes, or stroke are the most important predictors of LOS.

The LOS of patients with cardiac problems was the focus of the paper by Peyman Rezaei Hachesu et al., which is one of the few that employs machine learning techniques. [3] Patient data was retrieved over a 5 year period from a hospital in Iran that specializes in treating and researching cardiovascular conditions. 36 different attributes were included per row and three different models were run on the data: decision tree, neural network, and support vector machine. Out of the three, the support vector machine approach was the most accurate, with the diagnosis ICD-9 code, the diastoloc blood pressure and the age being the three most prominent input variables (highest relative weight). The ICD-9 code refers to International Statistical Classification of Diseases codes which provide an internationally standardized code per disease. Diastolic blood pressure refers to the blood pressure in the arteries between heart beats.

In a paper from Daniel Sessler et al., a retrospective review of a database was conducted to determine predictive factors for hospital stay and mortality. Analysis was done on a database from the Cleveland Clinic who had undergone noncardiac surgery within a five-year period along with measurements of mean arterial pressure (MAP), bispectral index (BIS), and minimum alveolar concentration (MAC). BIS refers to a meaure of the depth of anesthesia and MAC refers to the concentration of anethesic vapours used. Through logistic regression, it was found that a "triple-low value of MAP, BIS, and MAC were strongly correlated with an extended LOS [12]

In their paper, Andrew Kramer and Jack Zimmerman sought to create a predictive model for early identification of patients that were at risk for a prolonged ICU LOS [10]. Analysis was done on data collected on admission and on day 5 of a patients stay. Through a multivariable regression model, it was found that factors collected on day 5 had the greatest impact on LOS rather than those collected during admission. Some of these important factors include whether mechanical ventilation was required, PaO2:FiO2 ratio, assorted physiological components, and sedation.

The current literature demonstrates that multiple and complex factors can affect LOS and that these are not yet fully understood and that a comprehensive predictive model is currently lacking.

II. Datasets-Tools

In this section we will describe all the components of our experiment, giving brief descriptions of the dataset and the algorithms used. We will also discuss the mathematical and computer science foundation of the algorithms making the paper as self contained as possible.

A. MIMIC

A crucial component of this project is to be able to analyze the large amount of data contained in Electronic Medical Records (EMRs). Our dataset of choice was the MIMIC III database which houses over 50,000 actual records of people that have been admitted to ICU units between 2001 and 2012.

As one would expect, the MIMIC database contains data from anonymized EMRs with the data stored in a relational database format. Besides the actual raw data, tables of metadata are provided and an online query tool is available. The MIMIC dataset amongst other information covers the following types of information:

- Admissions
- Discharges: details of patient leaving hospital
- Transfers: records of the movement of patients between careunits and wards during their stay
- Caregivers: details of what type of staff cared for a patient during their hospital stay
- Prescriptions: medication information from the hospital computerized hospital order entry (CPOE) system
- Chart information: the patient's medical chart including such information as vital signs and ventilator settings
- Labevents: lab tests
- Diagnoses: various information on diagnoses from across multiple tables
- Procedures: various information on the procedures carried out on patients from across multiple tables

- Patients: demographic information on patients
- Patient notes: the patient notes recorded for each patient

In order for us to have immediate access to the raw data, a dedicated secure server was created at Florida Polytechnic University's supercomputer, where a copy of the entire database was downloaded.

B. Neural Networks

One of the most prominent tools for analyzing big data sets and generally data that comes from different modalities is Artificial Neural Networks. These algorithms emulate the "learn by example" technique that we use to understand a phenomenon.

At the core of a NN sits a directed graph of nodes and connectors, called neurons and synapses. Each layer of neurons is connected to the next one using a set of activation functions. The parameters of each of these functions are trained in parallel using a multivariate calculus optimization that minimizes an appropriate error function.

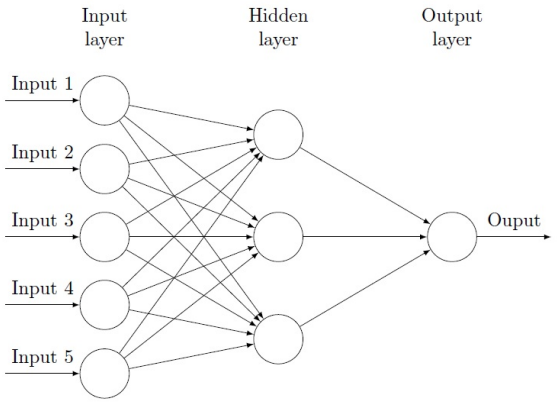

Fig. 1. A typical NN

Large annotated data sets have become increasingly available. The rise of the internet, the rapid increase in the use of electronic health records since legislative efforts in the mid-2000s, historical data, accurate computer simulations and other sources have led to further health-related datasets becoming available. In addition the field of neural networks (NN) has also become more associated with what has become known as deep learning. Neural networks require large data sets to train, especially when the input for a specific model can be broken down to many categories. However due to the rapid increase in the availability of digitized health data, neural networks have moved from a promising tool to potentially provide a powerful approach for research and health anlytics applications, especially when there is little or incomplete understanding of the underlying structure of a problem.

The input of these networks can be anything that is vectorizable. This makes them ideal in situations where the data set contains pictures, unstructured text and of course the traditional structured matrix with various features, numerical, Boolean etc. The flexibility of these algorithms allows us to combine the different modalities and get meaningful outcomes without having to discard unstructured information that may prove useful.

Although there are no clear measurements of accuracy, it has been observed that the predictability improves with the number of samples in the training set. Moreover, there is a threshold after which the network starts to be useful.

The output of a neural network could be binary, numerical, or even text and synthesis of images. It is agnostic to the types of input, but it is faster and more accurate when the input is structured numerical matrices. Implementations of the NNs in parallel systems are available in free online software. There are modifiable packages written in python, R, C++, and complete systems in software like matlab, SPSS, SAS and Torch.

C. Random Forests

A Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It is an *Ensemble Method* where multiple analytical models, called weak learners, that cover different parts of the training data are used together to form one powerful model, called a strong learner.

Random forests were first generally introduced by Tin Kam Ho 1995 [7]. Later in 2001 Breiman [2] properly introduced the concept by extending Ho's algorithm and the work of Amit and Geman [1].

The random forest starts with a standard machine learning technique called "decision trees" (the weak learners). In a decision tree, an input is entered at the top and as it traverses down the tree the data gets put into smaller and smaller sets or classes. The random forest goes a step beyond by combining multiple trees to produce the mode class, in the case of classification, or, in the case of regression, a mean prediction by averaging the results of the individual trees. In addition, random forests correct for decision trees' habit of overfitting to their training set [6].

More specifically a Random forest works as follows:

Given a data set with n observations and N variables. Fix a constant m with $m < N$.

Now:

1) Randomly sample n observations with replacement, this is called a *Boostrap Sample*.
2) Build a decision tree where at each node m randomly selected predictor variables are chosen on which to base the splitting decision. The best split on these m inputs is used to split the node.
3) Grow each tree to the largest extent possible without pruning.
4) Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

When a new input is entered into the system, it is run down all of the trees. In the case of a numerical variable, the

result may either be an average or weighted average of all of the terminal nodes that are reached.
However, if the input variable is categorical, then the result will be a voting majority.

Note that:

- With a large number of predictors, the eligible predictor set will be quite different from node to node.
- The greater the inter-tree correlation, the greater the random forest error rate, so one pressure on the model is to have the trees as uncorrelated as possible.
- As m goes down, both inter-tree correlation and the strength of individual trees go down. So some optimal value of m must be discovered.

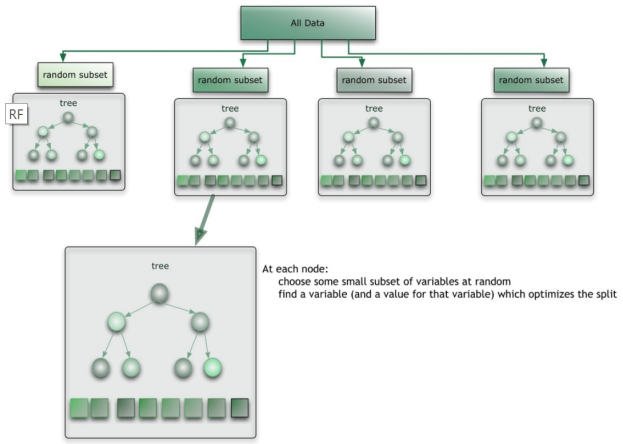

Fig. 2. A typical Random Forest

Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. One of their weaknesses is that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

III. Experiment

A. Experiment Description

Our main goal in this paper was to give an estimate of a patient's full Length Of Stay right after they exit the ICU unit. We wanted the input variables to be as general as possible and tried not to use specific tests or lab results. Specifically the question we wanted to answer was:

"Will the patient have a "long" stay at the hospital ($>$ 5 days) or short ($\leq$ 5 days) as determined at the point that they exit the ICU".

As we see in the following histogram (Fig.3), 5 is close to the mean, median and mode for the distribution of LOS, especially when we remove outliers of more than 20 days so it was a natural choice for a cut off point for considering short versus long stay.

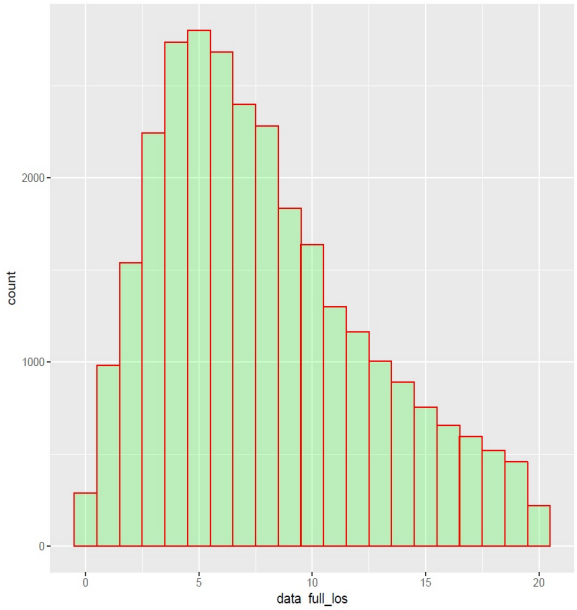

Fig. 3. Histogram for the Length Of Stay

As we mentioned before the database we used is the MIMIC III database and our input contains attributes from a combination of tables as described below:

1) **Admissions:** hadm_id (hospital admission ID), subject_id, admittime (time of admission), dischtime (time of discharge), ethnicity, admission_type, admission_location, insurance, religion, marital_status
2) **CPT Events:** costcenter, cpt_cd
3) **ICU Stays:** first_careunit, last_careunit, first_wardid, last_wardid, los
4) **Services:** prev_service, curr_service
5) **Patients:** gender, dob
6) **Procedures ICD:** seq_num, icd9_code (truncated to first 2 characters)
7) **Diagnoses ICD:** seq_num, icd9_code (truncated to first 3 characters)

The choice of these attributes was partly informed by running various automated prediction algorithms using the WEKA software [4]. The package, "autoweka", indicated that the best model on a small portion of our data set was Random Forest, using the aforementioned attributes. In addition to this an analysis of the semantics of the data and a review of the current literature to consider previously identified factors affecting LOS was carried out. Via this means the dataset for training the neural network was identified and refined.

For each admission in the MIMIC III database, a number of diagnoses are recorded, ordered from the diagnosis with seq_num equal to 1, up to the diagnosis with seq_num of n, where n is the maximum number of diagnoses recorded for that patient and that admission. Diagnosis ICD9 codes provide an encoding to designate diagnoses with three, four or five digits,

but for our purposes we truncated codes to three digits which has the effect of grouping specific diagnoses into groups of cognate conditions. For the diagnosis $icd9\_code$, we have used the $icd9\_code$ corresponding to seq_num equal to 1 which generally corresponds to the primary diagnosis or the diagnosis with the greatest significance to that patient for that hospital stay.

In relation to diagnosis seq_num the maximum or count of the sequence numbers is used, representing the total number of diagnoses for that patient for that admission.

Three variables were also created for our table: $full\_LOS$ which denotes the total LOS based on the period between discharge and admittance time and which is also the output variable to be predicted and age which determines the patient's age based on admittance time and dob. We also replaced the procedures sequence number and the procedures ICD code with the attribute "Procedures number" to reflect the total number of procedures performed on the patient during the stay at the ICU, without focusing on the specific type of those procedures.

To have a coherent data set we removed any row for which the patient's primary diagnosis ICD9 code started with a 'V' or 'E' and capped the total length of stay to 100 days. We chose that cut off point for our analysis to avoid skewed fit in our NNs due to the few spurious entries of very long LOS. To develop our predictive model, an R (https://www.r-project.org/) script was written that creates and trains an artificial neural network using the previously described table. The code operates as follows:

1) Several variables are established in order to create flexibility in the code. Specifically, one can adjust the total number of data points to use, the percentage of said data to use for training, the maximum LOS that should be included in the data, and the number of repetitions for the NN. This facilitates the implementation of various random forest configurations.
2) The data goes through a pre-processing stage. This includes reading in the data, selecting particular attributes for analysis, and cleaning up the data for incomplete data points and LOS longer than the determined maximum number of days.
3) The NN is created and trained using the scaled data. We are using the R package neuralnet [13]. The data is first randomly separated into training and testing sets and then scaled before being used by the NN. Our NN consists of 2 layers with 5 and 3 nodes in the corresponding layers. We came up with this schema after testing various configurations and choosing the one with the highest accuracy in the prediction.
4) After the NN is trained, testing is done in order to assess its accuracy. With the newly trained model, predictions are made on the testing set and stored in a variable. For our assessment, we wanted to assess the ability of the NN to predict a LOS of greater or less than 5 days. Accordingly, we sorted predictions into bins and assessed the accuracy using a confusion matrix, a standard representation for the performance of a classificaiton model. That accuracy is saved in a .csv and the confusion matrix is saved as an .rda.

In our final data set we had $31,018$ data points, each data point corresponding to a different patient. From it we created 100 subsets of $15,000$ points chosen at random. For each of this subsets we chose $13,500$ points (90%) and trained a NN with the process described above. The remaining $1,500$ points of each subset were used as test. This created 100 confusion matrices and 100 accuracy scores. The computations were done in parallel on the Florida Polytechnic University supercomputer. We describe our findings in detail in the next subsection.

B. Results

The box plot (Fig.4) and the accuracy statistical description table (Table I) below show that the neural network produced models that were accurate in predicting long vs short length of stay approximately 79% of the time. In particular, 50% of the models had accuracy scores between 78% and 79.8% with the median score being 79%. In addition, all 100 models had accuracy scores between 75.3% and 82.3% except for only two models with accuracy scores of 75% and 75.1%.

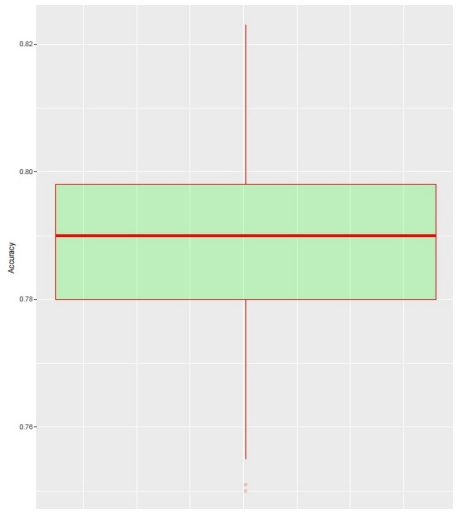

Fig. 4. Accuracy Scores Box Plot

Furthermore, looking at the histogram of the accuracy scores below (Fig.5), we notice that the accuracy scores are nearly normally distributed with a mean of 78.9% and standard deviation of 1.5%. Consequently, our 95% confidence interval is from 78.885% to 78.825%.

In comparison, a linear fit model on the entirety of the dataset yielded an accuracy of 57%. It is worth noting also, that the linear model did not make accurate predictions for all test cases where the length of stay was short, leading us to discard it immediately.

TABLE I
ACCURACY STATISTICAL DESCRIPTION

Statistic	Value
Number of Observations	100.0000
Minimum	0.750000
Maximum	0.823000
Range	0.073000
1st Quartile	0.780000
Median	0.790000
3rd Quartile	0.798000
Mean	0.788550
Standard Deviation	0.015140
Variance	0.000229
Mean Standard Error	0.001514
95% Confidence Interval.	0.003004

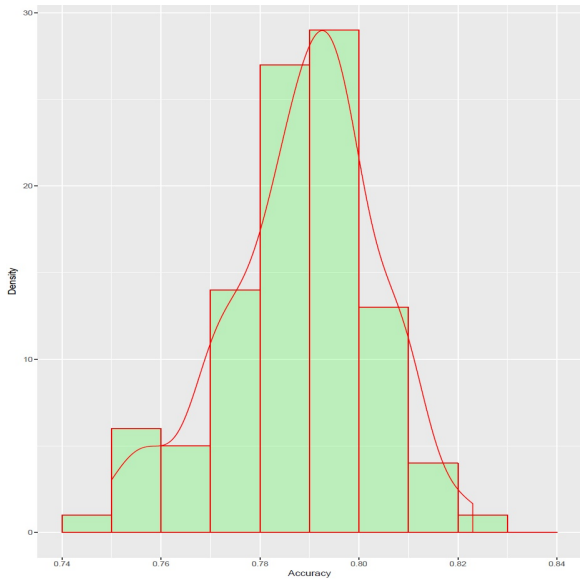

Fig. 5. Accuracy Scores Histogram

IV. DISCUSSION

The predictive model developed demonstrates good accuracy for predicting short versus long hospital stays particulary given that it is a model not constrained to or trained for a specific health condition or group of conditions as has been the case for previous length of stay NN results. The NN approach developed also did not make use of detailed time-stamped physiological patient data, care-giver observations, medication information, laboratory or imaging results, all sources of data that could have a potential predictive bearing on length of stay.

One of the benefits of this approach is that a relatively simple model in terms of inputs needed can be applied across all patients. In addition these inputs are known for all patients and via adminstrative processes rather than clinical data capture or tests. For example, many of the inputs are the demographics of a patient that are necessarily known at the point of admission: age,gender, marital status, ehnicity and religion. Similarly the ICU related inputs are known from transfer records such as the ICU care units in which a patient stayed.

The set of attributes inputted to the neural network are not varying per health condition of the patient. The implication of this is not just the general and condition-agnostic nature of the model, but the model is simple enough to be applied by hospital administrative support not requiring domain-specific clinician input to calculate.

One way in which the designation of the condition of a patient is utilized, but again not physiological data or tests specific to that condition, is via the use of the attribute containing diagnosis ICD9 code. The neural network currently only utilizes the ICD9 code of the primary diagnosis, although it could be extended to input the ICD9 codes of all concurrent diagnoses for the patient. These co-occuring diagnoses would provide more information on the specifics of the health of the patient but with each increasing sequence number, a generally decreasing level of priority or relevance of that diagnosis to the admission. It is of note that whilst using minimal diagnosis information, only the primary diagnosis, the neural network is still able to predict with good accuracy short versus long stay.

In addition to the ICD9 code of the primary diagnosis of each patient the count of the total number of diagnoses for that patient for that admission, is also used as an input. This is equivalent to the value of the maximum diagnosis sequence number for that patient for that admission. The higher this number for a given patient, the more health conditions or diagnoses the patient has. This has the potential to have some relationship to the state of their health and hence a relationship to length of stay.

Similarly procedures-related information is relatively lightly used. Procedure ICD9 codes denote the procedures performed on a patient in the hospital, and the order these are performed is desingated by the sequence numbers, with a sequence number of 1, indicating that is the first procedure carried out on that patient for that admission. The ICD9 codes for the procedures carried out are not currently input to the neural network. Rather just a count of the total number of procedures performed is currently inputted. The ICD9 code of the first occuring procedure and indeed the codes of all procedures occuring during that patients' admission may be expected to provide additional informaiton about the health status of the patient and potentially have a bearing on LOS.

Demographic inputs age, gender and marital status are used by the neural network, and such inputs have also been found to be factors affecting length of stay in previous condition-specific studies of LOS [3], [9].

The cost center input variable taken from the MIMIC III CPT_EVENTS table provides some limited amount of information on the nature of the patient's condition, having just two possible values ICU or RESP (respiratory).

The first_careunit and last_careunit inputs act to provide a proxy for information about the broad type of health coniditon and category of care of the patient. For example the possible values for careunits include such values as coronary care unit

(CCU), cardiac surgery recovery unit (CSRU), medical intensive care unit (MICU), neonatal intensive care unit (NICU), trauma/ surgical intensive care unit (TSICU). As such these careunit values can have a bearing on LOS.

Similarly the input curr_service serves as a proxy to provide information on the broad category of health condition. The service refers to the type of care team that is caring for the patient under which the patient was admitted and can have such values as dental (DENT), gynecological (GYN), general service for internal medicine (MED), psychiatirc (PSYCH) and cardiac surgery (CSURG) amongst a total of twenty possible values. The service can be a better indicator of the category of care or broad area of health condition than the ward that the patient is staying in, as in some cases a patient will be placed in a particular physical ward based upon bed availability even when it does not correspond with the health care team. For an entry in the MIMIC III SERVICES table, there will often be no value for prev_service and hence it is not as valuable - this is the case where the record reflects the patient's first admission to the hospital as is the case for most entries in the SERVICES table.

A strength of neural networks is that they allow prediction where the relationships between the input variables are complex and unknown. This makes the neural network well suited to the LOS prediction task. A recognized disadvantage of neural networks is that they require large amounts of data to train, however in the case of this work we have been able to utilize the MIMIC III database which provides a large data set with over 50,000 unique patients included and highly detailed medical data about each patient admission.

As such an approach based upon neural networks is well-suited to extension via the inclusion of additional patient data items. Whilst patients are present in an ICU ward, detailed data such as are captured and such data are present in the MIMIC III database. The current literature has yet to determine the inter-relationships between many of the factors affecting LOS in the general and condition-specific cases. As such neural networks provide a potentially benefical tool for LOS prediction for more complex patient data inputs.

V. Future Work

Future work includes most immediately inclusion of the full sets of diagnosis ICD9 codes rather than just the first code. This can potentially allow the effect of co-morbidities to be taken into account by the model. All have the potential to have a bearing on predicted LOS. It should be noted that beyond truncation of ICD9 diagnosis codes to three digits there are various other ways that the diagnoses can be appropriately grouped to cluster diagnoses with cognate characteristics within the same group. This development can also potentially take into account the relative importance of the different diagnoses based upon sequence number.

Inclusion of the first procedure ICD9 code, that is the ICD9 code of the first procedure carried out during an admission, may also prove beneficial. The nature of the procedure(s) carried out would be expected to provide further information pertaining to length of stay. Similarly the inclusion of all of the procedure ICD9 codes may be beneficial. An area of exploration is the relative importance of the various procedures based upon sequence number to affect LOS.

Further development involves the removal of ICU length of stay from the inputs, and predicting both the ICU length of stay and the hospital length of stay. This would have additional high clinical value as the cost per night for ICU-based care is typically multiples of the cost per night of a non-ICU ward bed, and so estimating ICU LOS has a per night greater implication for resource planning. In addition this can enable the point in time at which the predictive model can make the LOS prediction to be earlier. To be able to make a prediction of LOS earlier will also be increasingly desirable.

By limiting prediction to specific conditions, further results can be determined. This will support consideration of factors relevant to just specific health conditions and not all. Specific conditions groupings to be considered may be those most prevalently represented in the MIMIC III database which include:

- diseases of the circulatory system such as ischemic heart disease
- diseases of the digestive system
- trauma
- pulmonary diseases
- infectious and parasitic diseases
- neoplasms of digestive organs and intrathoracic organs

The includion of specific physiological data measurements, care-giver observations, lab and imaging test results and medication information, all provide the possibility of further predictive capabilities in both the general and condition-specific cases. Whilst in the ICU examples of types of patient data captured and recorded in MIMIC III include:

- vital signs,
- waveforms,
- trends,
- fluids

These pose the potential, due to the current availability of more detailed, large digitized data sets such as MIMIC III, to develop via machine learning approaches more sophisticated LOS predictive models than have been developed to-date.

VI. Conclusion

In this paper we explore the use of neural networks for predicting the total length of stay of a patient in hospital. Various patient and admission data are used to predict whether a stay will be long (> 5 days), or short (≤ 5 days), predicted at a point of time at which the patient is transferred out of the ICU unit. The system was trained on the large data set provided by the MIMIC III database. The predictive model performs with an accuracy, the precentage of long or short stays accurately predicted from the test set, of approximately 80%. This provides a more general LOS prediction system based on neural networks than has been previously reported.

All our codes are available at our gitlab repository. The dataset can be made available upon request through the MIMIC III website.

REFERENCES

[1] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[2] Leo Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[3] Peyman Rezaei Hachesu, Maryam Ahmadi, Somayyeh Alizadeh, and Farahnaz Sadoughi. Use of data mining techniques to determine and predict length of stay of cardiac patients. In *Healthcare Informatics Research*, pages 121–129. The Korean Society of Medical Informatics, June 2013. https://synapse.koreamed.org/search.php?where=aview&id=10.4258/hir.2013.19.2.121&code=1088HIR&vmode=FULL.

[4] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[5] Mohamad H. Hassoun. *Fundamentals of Artificial Neural Networks*. MIT Press, Cambridge, MA, USA, 1st edition, 1995.

[6] Allen Hatcher. *The Elements of Statistical Learning (2nd Ed.)*. Springer, 2008.

[7] Tin Kam Ho. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC*, pages 278–282, 1995.

[8] A. E. Johnson, T. J. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035, May 2016.

[9] Maria Kelly, Linda Sharp, Fiona Dwane, Tracy Kelleher, and Harry Comber. Factors predicting hospital length-of-stay and readmission after colorectal resection: a population-based study of elective and emergency admissions. *BMC Health Services Research*, 12(1):77, 2012.

[10] Andrew Kramer and Jack Zimmerman. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. In *BMC Medical Informatics and Decision Making*, pages 10–27. BioMed Central, May 2010. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-27.

[11] John Rapoport, Teres, Daniel MD, Zhao Yonggang, and Stanley Lemeshow. Length of stay data as a guide to hospital economic performance for icu patients. *Medical Care*, 41:386–397, 2003.

[12] Daniel Sessler, Jeffrey Sigl, Scott Kelley, Nassib Chamoun, Paul Manberg, Leif Saager, Andrea Kurz, and Scott Greenwald. Hospital stay and mortality are increased in patients having a "triple low" of low blood pressure, low bispectral index, and low minimum alveolar concentration of volatile anesthesia. In *The Journal of the American Society of Anesthesiologists*, pages 1195–1203. Web of Science, June 2012. http://anesthesiology.pubs.asahq.org/article.aspx?articleid=1933605.

[13] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[14] Ying Wang, Knut Stavem, Fredrik A Dahl, Sjur Humerfelt, and Torbjorn Haugen. Factors associated with a prolonged length of stay after acute exacerbation of chronic obstructive pulmonary disease (aecopd). In *International Journal of COPD*, pages 99–105. Dovepress, January 2014. https://pdfs.semanticscholar.org/5106/28c5594c21b145f0c012c6bc353f96fc71d0.pdf.