A Supervised Learning Approach for ICU Mortality Prediction based on Unstructured Electrocardiogram Text Reports

Gokul S Krishnan[1] and Sowmya Kamath S[2]

Department of Information Technology,
National Institute of Technology Karnataka
Surathkal, INDIA - 575025
[1]gsk1692@gmail.com [2]sowmyakamath@nitk.edu.in

Abstract. Extracting patient data documented in text-based clinical records into a structured form is a predominantly manual process, both time and cost-intensive. Moreover, structured patient records often fail to effectively capture the nuances of patient-specific observations noted in doctors' unstructured clinical notes and diagnostic reports. Automated techniques that utilize such unstructured text reports for modeling useful clinical information for supporting predictive analytics applications can thus be highly beneficial. In this paper, we propose a neural network based method for predicting mortality risk of ICU patients using unstructured Electrocardiogram (ECG) text reports. Word2Vec word embedding models were adopted for vectorizing and modeling textual features extracted from the patients' reports. An unsupervised data cleansing technique for identification and removal of anomalous data/special cases was designed for optimizing the patient data representation. Further, a neural network model based on Extreme Learning Machine architecture is proposed for mortality prediction. ECG text reports available in the MIMIC-III dataset were used for experimental validation. The proposed model when benchmarked against four standard ICU severity scoring methods, outperformed all by 10-13%, in terms of prediction accuracy.

Keywords: Unstructured Text Analysis, Healthcare Analytics, Clinical Decision Support Systems, Word2Vec, NLP, Machine Learning

1 Introduction

Identifying individuals who are at risk of death while admitted to hospital Intensive Care Units (ICUs) is a crucial challenge facing critical care professionals. Extensive and continuous clinical monitoring of high-risk patients is often required, which, given the limited availability of critical care personnel and equipment, is very expensive. Several mortality risk scoring systems are in use currently in ICUs that rely on certain patient-specific diagnostic and physiological factors identified by medical experts, extracted from structured health records (EHRs), to calculate mortality risk. However, studies have reported that their

performance in actual prediction is quite low when compared to more recent non-parametric models based on data mining and Machine Learning (ML) [18, 4]. As structured EHRs are put together manually with extensive human effort, a lot of context information contained in clinician's notes might be lost [17]. Another significant issue is the limited adoption rate of structured EHRs in developing countries, thus necessitating the use of alternate methods to obtain patient-specific information [22]. Most existing Clinical Decision Support System (CDSS) applications [4, 10, 18] depend on the availability of clinical data in the form of structured EHRs. However, in developing countries, clinical experts and caregivers still rely on unstructured clinical text notes for decision making. Unstructured clinical notes contain abundant information on patients' health conditions, physiological values, diagnoses and treatments, are yet to be explored for predictive analytics applications like mortality risk prediction and disease prediction. Such unstructured clinical text represent a significant volume of clinical data, which has remained largely unexploited for building predictive analysis models. Big data analytics and ML can help in developing better CDSSs with significant man-hour and medical resource savings [2].

Traditional methods for computing ICU mortality comprise of the standard severity scoring systems currently in use in hospitals. APACHE (Acute Physiology And Chronic Health Evaluation) [11] and SAPS (Simplified Acute Physiological Score) [6], along with their different variants; SOFA (Sequential Organ Failure Assessment) [21] and OASIS (Oxford Acute Severity of Illness Score) [8] are popular severity scoring models used for computing an ICU patient's mortality score using their clinical data. Other approaches [16, 5, 10, 18, 4] focus on application of ML techniques like decision trees, neural networks and logistic regression to structured EHR data for predicting mortality scores. Free text clinical notes written by medical personnel possess a significant volume of patient-specific knowledge, that is expressed in natural language. Several researchers proposed methods [23, 14, 19, 1] for making use of such data for various purposes like data management, patient record classification and event prediction using ML, Hidden Markov Models, genetic algorithm and several other natural language processing (NLP) and data mining techniques.

In this paper, an ICU Mortality prediction model that utilizes patients' unstructured Electrocardiogram (ECG) text reports is proposed. We adopt the Word2Vec word embedding models for vectorizing and modeling the syntactic and semantic textual features extracted from these reports. An unsupervised data cleansing technique designed for identifying and removing anomalous data and special cases is used for optimizing the data representation. Further, a neural network architecture called Extreme Learning Machine (ELM) is trained on the ECG data for mortality risk prediction. The proposed model when benchmarked against existing traditional ICU severity scoring methods, SAPS-II, SOFA, OASIS and APS-III, achieved an improved accuracy of 10-13%. The remainder of this paper is organized as follows: In Section 2, we describe the process of designing the proposed prediction model. Section 3 presents results of the experimental validation, followed by conclusion and references.

2 Proposed ICU Mortality Prediction Model

The methodology adopted for the design of the proposed mortality prediction model is composed of several processes, which are depicted in Fig. 1.

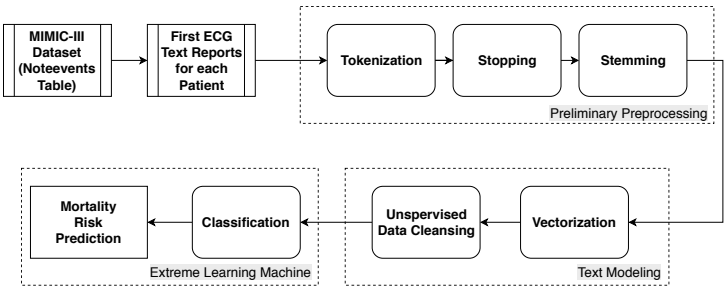

Fig. 1: Proposed methodology

2.1 Dataset & Cohort Selection

For the proposed model, unstructured text data from an open and standard dataset called MIMIC-III [9] was used. MIMIC-III (Medical Information Mart for Intensive Care III) consists of deidentified health data of 46,520 critical care patients. Clinical text records of these patients are extracted from the 'noteevents' table in the MIMIC-III dataset, from which only the ECG text reports are selected. Currently, we have considered only the first ECG report of each patient, as this is required to predict patients' mortality risk with the earliest detected condition, thereby predicting risk earlier. Next, the mortality labels of each patient are extracted from the 'patients' table in the dataset and are assigned to corresponding ECG reports of each patient. This set, now containing the first ECG text reports of 34,159 patients and the corresponding mortality labels, is used for the next phase (Details of ECG text corpus summarized in Table 1a).

2.2 Preliminary Preprocessing

In the next phase, the ECG text corpus is subjected to a NLP pipeline consisting of tokenization, stopping and stemming. During tokenization, the clinical natural language text is split into smaller units called tokens. Generated tokens are filtered to remove unimportant terms (stop words) and finally, stemming is performed on the remaining tokens for suffix stripping. After the initial preprocessing, the tokens are next processed for modeling any latent clinical concepts effectively, during the Text Modeling phase.

2.3 Text Modeling

Thee Text Modeling phase consists of two additional levels of processing - Vectorization and Unsupervised Data Cleansing, which are discussed in detail next.

Table 1: Dataset Statistics

(a) ECG Text Corpus Statistics

Feature	*Total Number*
Reports	34159
Sentences	108417
Total Words	802902
Unique Words	33748

(b) Statistics of the selected patient cohorts

Set	*Total*	*Alive*	*Expired*
Initial ECG Text reports	34159	30464	3695
Cluster C_1	22974	20372	2602
Cluster C_2	11185	10092	1093
Final Cohort	21465	20372	1093
Training & Test sets	10155	8068	2087
Validation set	2539	2024	515

Vectorization : NLP techniques are critically important in a prediction system based on unstructured data, for generating machine processable representations of the underlying text corpus. Traditional rule and dictionary based NLP techniques, though perform well for certain applications, are not automated and require significant manual effort in tailoring them for various domains. Recent trends in ML and Deep Learning models and their usage in addition to traditional NLP techniques provide a good avenue for exploiting their performance for improved prediction. However, the effectiveness and performance of such models depend heavily on the optimized vector representations of the underlying text corpus. Several approaches have been developed for creating meaningful vector representations from text corpus, the prominent ones being Document Term Frequency vectorization and Term frequency-Inverse document frequency (Tf-Idf) Vectorization [20]. Word2Vec [15], a word embedding model, is an effective approach for generating semantic word embeddings (features) from unstructured text corpus. The generated vectors may be of several hundred dimensions, where unique terms in the text corpus are represented as a vector in the feature space such that corpus terms of similar context are closer to each other [15]. For modeling such latent concepts in the ECG text report corpus, we employed Word2Vec to generate a word embeddings matrix, which consists of the syntactic and semantic textual features obtained from the unstructured ECG corpus. The skip-gram model of Word2Vec was chosen over Continuous Bag-Of-Words (CBOW), due to its effectiveness with infrequent words and also as the order of words is important in the case of clinical reports [15]. We used a standard dimension size of 100, i.e., each ECG report is represented using a 1 x 100 vector, thus resulting in a final matrix of dimension 34159 x 100, each row representing the latent concepts in the ECG report of a specific patient.

Unsupervised Data Cleansing: The vectorized ECG text corpus data is next subjected to an additional process of data cleansing, for identifying special case data points and conflicting records. For this, K-Means Clustering was applied on the vectorized data to cluster the data into two clusters (k=2, as the proposed prediction model is a two-class prediction, 'alive' and 'expired' patients) after which a significant overlap was observed in the two clusters. Cluster C_1 contained

records of 20372 alive and 2602 expired patients while cluster C_2 had 10092 alive and 1093 expired patients. As a significant number of the data points representing 'alive' patients were in cluster C_1, we derived a reduced patient cohort that consists of all 'alive' patients from cluster C_1 and all 'expired' patients from cluster C_2, which were then considered for building the prediction model. The remaining patient data points exhibited anomalies due to existence of patients who might have expired due to causes not related to heart. Further examination of these special or conflicting cases revealed a requirement for considerable auxiliary analysis, therefore, we intend to consider them as part of our future work. In summary, the new patient cohort now consisted of 20372 alive and 1093 expired patients (tabulated in Table 1b).

2.4 Linguistics driven Prediction Model

The patient cohort obtained after the data cleansing process is now considered for building the prediction model. Towards this, we designed a neural network model that is based on a fast learning architecture called Extreme Learning Machine (ELM) [7]. ELM is a single hidden layer Feedforward Neural Networks (SLFNN) where the parameters that fire the hidden layer neurons don't require tuning [7]. The hidden nodes used in ELM fire randomly and learning can be carried out without any iterative tuning. Essentially, the weight between the hidden and output layers of the neural network is the only entity that needs to be learned, thus resulting in an extremely fast learning model. Different implementations of ELMs have been used for tasks like supervised and unsupervised learning, feature learning etc, but to the best of our knowledge, ELMs have not been applied to unstructured clinical text based prediction models. In this SLFNN architecture, we set the number of nodes in the input layer to 100 as the feature vectors obtained after Word2Vec modeling are of similar dimensions. The hidden layer consists of 50 nodes and a single node is used at the output layer, to generate the predicted mortality risk of a patient. Rectified Linear Unit (ReLU) activation function was used in the layers of the proposed ELM architecture as it is a step function and works well for binary classification. During training, the weights between the hidden and output layers are iteratively learned and optimized. Finally, the patient-specific mortality prediction is obtained at the output layer.

3 Experimental Results

For validating the proposed prediction model, an extensive benchmarking exercise was carried out. The experiments were performed using a server running Ubuntu Server OS with 56 cores of Intel Xeon processors, 128GB RAM, 3TB Hard Drive and two NVIDIA Tesla M40 GPUs. The patient cohort is split into training, test and validation sets (as shown in Table 1b). The vectorized feature vectors and the respective mortality labels in the training dataset are used for training the ELM model. We used 10-fold cross validation with a training to test data ratio of 75:25. Standard metrics like Accuracy, Precision, Recall, F-score and AUROC (Area under Receiver Operating Characteristic) were used for

performance evaluation of the proposed model. Additionally, Matthews Correlation Coefficient (MCC) was also used as a metric, as it takes into account true positives, false positives and false negatives, therefore, regarded as a balanced measure even in presence of class imbalance [3]. We also benchmarked the performance of the proposed prediction model against well established, traditional parametric severity scoring methods. Four popular scoring systems - SAPS-II, SOFA, APS-III and OASIS, were chosen for this comparison. We implemented and generated the respective scores for each patient in the validation set. For SAPS-II, the mortality probability was calculated as per the process proposed by Le et al [13]. In case of SOFA, the mortality prediction of each patient can be obtained by regressing the mortality on the SOFA score using a main-term logistic regression model similar to Pirracchio et al [18], whereas for APACHE-III (APS III), it is calculated for each patient as per Knaus et al' method [12]. The mortality probability for each patient as per OASIS scoring system is given by the in-hospital mortality score calculation defined by Johnson et al [8]. A classification threshold of 0.5 was considered for SAPS-II, APS-III and OASIS.

Table 2: Benchmarking ELM model against Traditional severity methods SAPS-II, SOFA, OASIS and APS-III

Models	*Accuracy*	*Precision*	*Recall*	*F-Score*	*AUROC*	*MCC*
ELM *(Proposed)*	0.98	0.98	0.98	0.98	0.99	0.84
SAPS-II	0.86	0.87	0.86	0.86	0.80	0.34
SOFA	0.88	0.86	0.88	0.85	0.73	0.22
APS-III	0.89	0.86	0.89	0.86	0.79	0.26
OASIS	0.88	0.86	0.89	0.86	0.77	0.26

The validation patient data is fed to the trained model for prediction and its performance was compared to that of traditional scoring methods. The results are tabulated in Table 2, where, it is apparent that the proposed model achieved a significant improvement in performance over all four traditional scores. The proposed model predicted high mortality risk (label 1) correctly for most patients belonging to 'expired' class, which is a desirable outcome expected out of this CDSS, which is also evident in the high precision values achieved. AUROC, F-Score and MCC are very relevant metrics for this experiment as the data exhibits class imbalance (number of patients in 'alive' class much greater those in 'expired' class; see Table 1b). MCC, which measures the correlation between the actual and predicted binary classifications, ranges between -1 and +1, where +1 represents perfect prediction, 0 indicates random prediction and -1 indicates total disagreement between actual and predicted values. The high values of F-Score and MCC for the proposed model in contrast to the others, indicates that, regardless of class imbalance in the data, the proposed model was able to achieve a good quality classification for both alive (0) and expired (1) labels. The plot of Receiver Operating Characteristic (ROC) curves generated for all models considered for comparison is shown in Figure 2. Again, it is to be noted that the proposed model showed a substantial improvement of nearly 19% in AUROC in comparison to the best performing traditional model, SAPS-II.

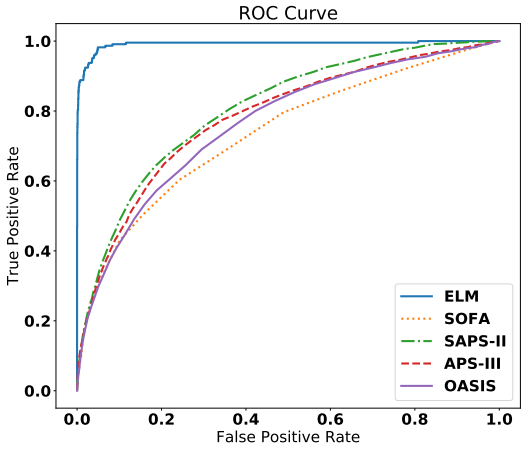

Fig. 2: Comparison of AUROC performance of the various models

4 Conclusion and Future Work

In this paper, a CDSS model for ICU mortality prediction from unstructured ECG text reports was presented. Word2Vec was used to model the unstructured text corpus to represent patient-specific clinical data. An unsupervised data cleansing process was used to handle conflicting data or special cases which represent anomalous data. A neural network architecture based on Extreme Learning Machines was used to design the proposed model. When benchmarked against popular standard severity scoring systems, the proposed model significantly outperformed them by 10-13%.

This work is part of an ongoing project with an objective of effectively using unstructured clinical text reports such as nursing/diagnostic notes and other lab test reports for improved decision-making in real-world hospital settings. Currently, our focus is using only a certain subset of the data for training and validation, however, in future, we intend to develop a prediction model that can perform well even in the presence of anomalous data. Further, we also intend to explore the suitability of deep learning architectures for unsupervised feature modeling for optimal corpus representation and improved prediction accuracy.

Acknowledgement

We gratefully acknowledge the use of the facilities at the Department of Information Technology, NITK Surathkal, funded by Govt. of India's DST-SERB Early Career Research Grant (ECR/2017/001056) to the second author.

References

1. Barak-Corren, Yuval, et al.: Predicting suicidal behavior from longitudinal electronic health records. American journal of psychiatry 174(2), 154–162 (2016)

2. Belle, Ashwin, et al.: Big data analytics in healthcare. BioMed research international 2015 (2015)
3. Boughorbel, Sabri, et al.: Optimal classifier for imbalanced data using matthews correlation coefficient metric. PloS one 12(6), e0177678 (2017)
4. Calvert, et al.: Using EHR collected clinical variables to predict medical intensive care unit mortality. Annals of Medicine and Surgery 11, 52–57 (2016)
5. Clermont, et al.: Predicting hospital mortality for patients in the ICU: a comparison of artificial neural networks with logistic regression models. Critical care medicine 29(2), 291–296 (2001)
6. Gall, L., et al.: A simplified acute physiology score for ICU patients. Critical care medicine 12(11), 975–977 (1984)
7. Huang, Gao, et al.: Trends in extreme learning machines: A review. Neural Networks 61, 32–48 (2015)
8. Johnson, et al.: A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. Critical care medicine 41(7), 1711–1718 (2013)
9. Johnson, et al.: MIMIC-III, a freely accessible critical care database. Scientific data 3 (2016)
10. Kim, et al.: A comparison of ICU mortality prediction models through the use of data mining techniques. Healthcare informatics research 17(4), 232–243 (2011)
11. Knaus, et al.: Apache-a physiologically based classification system. Critical care medicine 9(8), 591–597 (1981)
12. Knaus, et al.: The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100(6), 1619–1636 (1991)
13. Le Gall, J., et al.: A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. Jama 270(24), 2957–2963 (1993)
14. Marafino, Davies, et al.: N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. JAMIA 21(5), 871–875 (2014)
15. Mikolov, Chen, et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Nimgaonkar, et al.: Prediction of mortality in an indian icu. Intensive care medicine 30(2), 248–253 (2004)
17. Omalley, Grossman, et al.: Are electronic medical records helpful for care coordination? experiences of physician practices. Journal of general internal medicine 25(3), 177–185 (2010)
18. Pirracchio, et al.: Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): a population-based study. The Lancet Respiratory Medicine 3(1), 42–52 (2015)
19. Poulin, Shiner, et al.: Predicting the risk of suicide by analyzing the text of clinical notes. PloS one 9(1), e85733 (2014)
20. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513–523 (1988)
21. Vincent, et al.: The sofa score to describe organ dysfunction/failure. Intensive care medicine 22(7), 707–710 (1996)
22. Williams, F., Boren, S.: The role of the electronic medical record (emr) in care delivery development in developing countries: a systematic review. Journal of Innovation in Health Informatics 16(2), 139–145 (2008)
23. Yi, K., Beheshti, J.: A hidden markov model-based text classification of medical documents. Journal of Information Science 35(1), 67–81 (2009)