Predicting Success of Motivational Interviews with Recurrent Neural Network and Probabilistic Model for Patient-Provider Communication Sequences

Mehedi Hasan, BS[1*], Alexander Kotov, PhD[1*], April Idalski Carcone, PhD[2], Ming Dong, PhD[1], Sylvie Naar, PhD[2]
[1]Department of Computer Science, [2]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract *The problem of analyzing temporally ordered observation sequences to make predictions related to the outcome of the processes that generated these sequences arises in many domains of healthcare informatics. In this paper, we focus on patient-provider communication sequences in the context of clinical interviews and propose Recurrent Neural Network (RNN) with two baseline methods Markov chain and Hidden Markov Model (HMM), for predicting the likelihood of eliciting a particular type of patient behavioral response based on an observed sequence of patient-provider exchanges. Our method achieved 70.03%, 48.65%, 88.53% F1-score for under-sampled sequences and 78.96%, 77.06%, 87.16% F1-score for over-sampled sequences in predicting the outcome of motivational interviews with obese adolescents using Markov Chain, HMM, and RNN, respectively. The proposed method can be used to automatically identify the most effective communication strategies in motivational interviews, which significantly decreases the effort required to develop effective interventions to address many public health conditions.*

Introduction

Data in the form of temporally ordered sequences of discrete or continuous observations (e.g., symbolic sequences such as notes in patient EHR, diagnostic codes, protein or DNA sequences or continuous time series, such as ECG measurements) arise in various domains of health informatics. The order of observations in a sequence is challenging to capture using features, which makes sequence classification a more challenging task than traditional classification. In general, sequence classification methods can be divided into three categories: feature-based, distance metric-based and model-based method. Feature-based classification methods transform a sequence into a feature vector and apply a standard supervised learning algorithm, such as support vector machine[1] or decision tree[2]. Shapelet[3] and pattern[4,5] based techniques as well as hierarchical approaches[6] have been proposed instead of standard classifiers as well. Distance-based methods measure the similarity between sequences to determine the quality of classification. The most commonly used distance function is Euclidian distance[7] with Dynamic Time Wrapping[8] used for more flexible matching in time series data. The third type of sequence classification methods, first create a probabilistic model of a sequence, such as Hidden Markov Model[9] (HMM).

Sequence classification has a broad range of real word applications from genomics and health informatics to finance and anomaly detection. In genomics research, sequence classification is widely used to classify protein and text sequence data[10] and detect the function of new proteins[11]. In health informatics, ECG measurements are considered as multi-dimensional time series and are used to classify individuals as healthy or having a heart disease[12]. Sequence classification is also used for anomaly detection such as abnormal access to systems[13] and malware[14].

In this paper, we direct our focus towards patient-provider communication sequences in the context of clinical interviews. Specifically, we focus on the transcripts of Motivational Interviews (MI) with obese adolescents and their caregivers. Childhood obesity is a serious public health concern in the United States and worldwide. Recent estimates indicate that approximately one third (31.8%) of US children age 2-19 years are overweight and 16.9% are obese[15]. Adolescents who are obese likely continue to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis[16]. One approach to effective obesity intervention is Motivational Interviewing, an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change. The goal of MI counseling session is to encourage patients to explore their own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements, referred to as "change talk" (or CT), consistently predict actual behavior change that can be sustained for as long as 34 months after an interview. However, the ability of counselors to consistently elicit this type of patient communication requires knowledge of effective commu-

*Authors provided equal contribution.

nication strategies for a variety of patients, which can only be obtained through analysis of a large number of annotated interviews. Since manual annotation and examination of transcripts is a very time-consuming process, tailoring of MI interventions to particular populations can take years. Therefore, there is a need for informatics-based methods to facilitate the development of effective MI-based interventions, in general, and theoretically-grounded computational models to explore the mechanisms of MI's efficacy, in particular.

Our research addresses this problem from multiple angles. In our previous work[17,18], we explored several machine learning methods for automatic annotation of clinical interview fragments with a specialized codebook containing a large number of patient and provider behavior codes[19]. In this work, we propose probabilistic methods to identify patient-provider communication sequences that are likely to elicit the desired patient behavioral response (i.e. change talk or commitment language) and to dynamically estimate the likelihood of observing this desired response at any point during a clinical interview based on all coded previous patient-provider communication exchanges in the same interview. While there have been some previous qualitative studies of patient-provider dialog in a clinical setting[20], there have been no previous work on computational modeling of annotated patient-provider communication (PPC) exchanges and predicting a desired patient behavior in a context of motivational interviews.

Methods

Data collection

The experimental dataset for this work was constructed from the transcripts of 106 motivational interviews, which include a total of 41,764 segmented and annotated utterances. Each transcript consists of a counselor-adolescent and a counselor-caregiver session. Since the ultimate goal of a motivational interview is to increase the desire and ability of adolescents for a targeted behavior change, we considered only communication sequences from counselor-adolescent sessions and disregarded all communication sequences from counselor-caregiver sessions. The utterances were annotated based on MYSCOPE codebook[19], which are grouped into the adolescent and counselor code groups. Utterances were divided into successful and unsuccessful communication sequences. Successful communication sequences result in positive change talk and commitment language (a special class of change talk) statements by an adolescent, while unsuccessful sequences are the ones that result in negative change talk or commitment language and the sequences, in which no change talk or commitment language statements occur. Out of 4169 observed sequences, 3427 were positive and 742 were negative. For each of the probabilistic models (Markov chain and HMM), one model was trained using successful sequences and one model was trained using unsuccessful sequences. Statistics of experimental dataset are presented in Table 1 and a fragment of an adolescent session transcript is presented in Table 2.

Table 1: Statistics of experimental dataset. Sequence length is the number of behavior codes in it.

Sequence Type	# of sequences	Ratio	Average length
Successful sequences	3427	82.20%	10.05
Unsuccessful sequences	742	17.80%	9.89

As can be seen in Table 1, our data set is imbalanced and predominately composed of successful sequences with only a small percentage (17.80%) of unsuccessful sequences. Usually, predictive accuracy misrepresents the performance of an employed algorithm for an imbalanced data. Because simply predict the majority class (successful) would provide a predictive accuracy 82.20%. Therefore, it is important to handle imbalance data properly. Many solutions are proposed to deal with the imbalanced data sets, which can be divided into two categories: data and algorithmic levels. In this study, we utilized under and over sampling methods at data processing levels for balancing the adolescent obesity data set. Synthetic Minority Over-sampling (SMOTE) method is the most widely used method for over-sampling an imbalanced data set, in which new synthetic examples were generated from minority class[21]. In this study, we generated synthetic examples along the borderline between minority examples and their selected nearest neighbors[22]. On the other hand, under sampling method under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a K-Means algorithm. Annotation column in Table 2 shows the sequence of behavior codes from top to bottom, where counselor starts with an open-ended question and gets positive feedback at the end.

Table 2: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent.

Annotation	Description	Speaker	Text
331	Open-ended question, elicit change talk positive	Counselor	do you feel like making healthier choices for your snacks and your meals is something you would be able to do? mm-hmm meaning is that food available for you?
117	Low Uptake, positive	Adolescent	Yes
301	Structure Session	Counselor	okay and thats an important thing for us to think about cause i would not want to help you come up with a plan that you would not be able to do without somebody else help so the last part of your plan is how somebody could be supportive to you meaning how they can help you be successful and so we should choose somebody who you feel like is around often enough
112	Change Talk positive	Adolescent	my um aunt
301	Structure Session	Counselor	okay so lets stick something my aunt can do
112	Change Talk positive	Adolescent	she could when i am doing when i am eating something that i should i could not be eating but so i can choose something healthy she could tell me not to eat it
309	Affirm, low	Counselor	okay that sounds like a really great suggestion

Prediction method

Generally, a sequence is a temporally ordered set of events. In this study, an event is a behavior code that also has a symbolic representation, such as 117 $LUP+$ (low uptake, positive), 331 $OQ-ECT+$ (open-ended question, elicit change talk positive), etc. Given a sequence of behavior codes $S_i = \{c_1, c_2, ..., c_n\}$ representing patient-provider communication exchanges during some part of a motivational interview, the task of predicting interview success can be viewed as sequence classification. Given a set of class labels $L = \{l_1, l_2, ..., l_m\}$ (in our case, the labels are "successful" and "unsuccessful" motivational interview), a sequence classifier C learns a function $S_i \rightarrow l_i, l_i \in L$ that maps a sequence S_i into a class label $l_i \in L$.

Our proposed baseline prediction method consists of two steps. In the first step, we model successful and unsuccessful patient-provider interactions using first- and second-order Markov Chain (MC) and Hidden Markov Model (HMM), which are popular probabilistic models for discrete observation sequences with finite vocabulary. In the second step, we classified each test sequence based on the maximum likelihood of generating that sequence from each model. Although HMM was originally developed for speech recognition[9], it is one of the most widely used methods for sequence modeling[23–27]. The latest advances in deep learning technologis shows that deep learning, in particular RNNs, provide better results than conventional machine learning methods for the task of sequence classification. In our experiment, we also employed two especial variant of RNN models: long short-term memory (LSTM) and gated recurrent unit (GRU).

Markov Chain (MC) is a stochastic model for randomly changing systems, which assumes that the next state of a system only depends on its current state and not on its past states (Markov property). Generally, this assumption enables reasoning and computation with the models that would otherwise be intractable. For the sequential analysis, we built two Markov models M and $\overline{M}$ summarizing provider strategies and patient responses in case of successful (M) and unsuccessful ($\overline{M}$) motivational interviews. A Markov model M can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, AMB-, BLT, LUP+, LUP-, HUP-W, ...\}$ is a set of vertices, consisting of adolescent and counselor MI behavior codes;
- $E \subseteq V \times V$ is a set of edges corresponding to posssible transitions from one MI behavior code to the other in a sequence;

- $p_M : E \rightarrow [0...1]$ is a function that assigns probability $p(c_i|c_j)$ to an edge between the MI behavior codes c_i and c_j based on maximum likelihood estimator:

$$P_M(c_j|c_i) = \frac{n_{c_i,c_j}}{n_{c_i}} \tag{1}$$

where n_{c_i,c_j} and n_{c_i} are the number of times a transition between the MI behavior codes c_i and c_j and the code c_i has been observed in the training data. Given a Markov model M (such that $S \subseteq V$), the probability that a sequence of MI behavior codes $S = \{C_1, ..., C_N\}$ has been generated from a Markov model M is:

$$P_M(S) = \prod_{i=2}^{N} p_M(c_i|c_1, \ldots, c_{i-1}) = \prod_{i=2}^{N} p_M(c_i|c_{i-1}) \tag{2}$$

In the second step, we quantify the likelihood of success of a given motivational interview at a certain time point given a sequence of MI behavior codes S observed prior to that point using the following formula:

$$p(S \rightarrow CML+) = \log\left(\frac{P_M(S)}{P_{\overline{M}}(S)}\right) = \sum_{i=2}^{N} \log p_M(c_i|c_{i-1}) - \sum_{i=2}^{N} \log p_{\overline{M}}(c_i|c_{i-1}) \tag{3}$$

If $p(S \rightarrow CML+) > 0$, the interview is predicted to result in positive change talk or commitment language.

The Markov model we have discussed so far is referred to as the first-order MC, since it only considers immediately preceding behavior code when computing the state transition probabilities. A generalization of the first-order MC is the k^{th} order MC, in which the transition probability to a particular state (symbol) x_i is computed by looking at the k preceding states (symbols). Thus, k^{th} order Markov chain will have N^k states each associated with a sequence of k symbols. In our experiment, we also used second order Markov model. Hence, our model has N^2 states with transition probability matrix of size $N^2 \times N$.

Hidden Markov Model (HMM) is a powerful tool for statistical modeling of processes varying in time. HMMs are widely used for sequence analysis because of their ability to incorporate dependencies among elements in sequence. HMM can be considered as a doubly embedded stochastic process with a process that is not observable (hidden process) and can only be observed through another stochastic process (observable process) that produces a sequence of observations. An HMM can be fully specified by the following quintuple: $\lambda = (N, M, A, B, \pi)$

- N is a number of hidden states in the model
- M is a number of distinct observations symbols per state, i.e. the discrete vocabulary size
- A is an $N \times N$ state transition probability distribution matrix $A = \{a_{ij}\}$
- B is an $N \times M$ matrix $B = \{b_j(k)\}$ with observation symbol probability distribution for each state
- π is the initial state distribution vector $\pi = \{\pi_i\}$

We exclude the structure parameters M and N to designate HMM using a compact notation: $\lambda = (A, B, \pi)$. The key difference between HMM and MC is that HMM requires specifying the number of hidden states as a model parameter and then the model deduces a sequence of hidden states that best explains the observations along with state transition probabilities and distributions of observation symbols for each state. The Baum-Welch algorithm was used to estimate the parameters of HMMs for successful and unsuccessful interviews using the corresponding training set, while the Viterbi algorithm was used to determine the most likely sequence of hidden states for a given sequence of observations. After assignment of hidden states, the log-likelihood of successful outcome can be estimated using Eq. 3.

Recurrent Neural Networks (RNN) is a type of Neural Network, which can send feedback from the current hidden state to the hidden state of the next time step. In this way, RNN can capture long-term dependencies for predicting the

future events, which is the main advantage of RNN over feedforward Neural Network, MC, and HMM. The capability of remembering past event is very useful in motivational interviews where a behavior observed at some points in the interview is very informative for the future behaviors that will be observed. However, it was observed that RNN fails to capture long-term dependencies due to vanishing gradient problem[28]. In order to mitigate this problem, Hochreiter et al.[29] proposed a special kind of RNN called Long Short Term Memory networks or simply LSTM. Later, the model was enhanced by including forget gates and the tanh activation function[30]. There are several variants of LSTM model, where a dramatic variation on the LSTM is the Gated Recurrent Unit or simply GRU[31]. GRUs are simpler than LSTM units and experimentally outperform than other models in many cases. We reiterate the mathematical formulation of GRU defined by Chung et al.[32] as follows (equation 4 to 7):

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{5}$$

$$\tilde{h}_t = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{6}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{7}$$

$$total\ loss = \alpha \cdot \frac{1}{T} \sum_{t=1}^{T} loss(\bar{y}^{(t)}, y^{(t)}) + (1 - \alpha) \cdot loss(\bar{y}^{(T)}, y^{(T)}) \tag{8}$$

The architecture of proposed Recurrent Neural Network is shown in Figure **??**. We used one hidden LSTM or GRU layer of 20 nodes after embedding the given myscope codes from input sequence. Then, we applied a softmax layer at each time step for predicting the target. This target replication strategy was used in several studies[33,34] after introduced by Lee et al.[35] in 2015 for convolutional neural networks. As it can be seen from equation (8), our loss is the combination of final loss and the mean of the losses over all sequence steps, where T is the total number of sequence step, $\bar{y}$ is the output at step t, and $\alpha \in [0, 1]$ is a hyperparameter that determine the relative importance of intermediate outputs. LSTM and GRU are trained on 70% of the data and validated on 20%. The remaining 10% of the data is used as a test set for reporting the performance of the model. Our models contain several hyperparameters such as embedding dimension, number of hidden units, learning rate, batch size, etc., which were determined by the validation set. We implemented our models in Tensorflow with adam optimizer and early stopping based on the validation loss.

Evaluation metrics

Performance of the proposed method was evaluated in terms of precision, recall, and F-measure using 10 folds cross-validation and weighted macro-averaging of these metrics over the folds. A true positive (TP) was counted when the method correctly classified a sequence into its actual class; a false positive (FP) was counted for a class when the method incorrectly classified a sequence into that class; a false negative (FN) for an actual class of sequence was counted when the method incorrectly classified the sequence into other class. Precision of a class was defined as the ratio of the numbers of correctly classified sequences and all sequences identified as belonging to that particular class by the classifier (i.e Precision = TP / (TP + FP)). Recall of a class was defined as the ratio between the numbers of correctly classified sequences and all sequences of that particular class in the gold standard (i.e. Recall = TP / (TP + FN)). F-measure is computed as the harmonic mean of precision and recall (i.e. 2 x Precision x Recall / (Precision + Recall)). However, Accuracy is simply computed as the ratio of correctly classified sequences and total number of sequences.

Results

Experimental evaluation of the proposed method is conducted on both under- and over-sampled sequences.

Predictive performance for under-sampled PPC code sequences

Predictive performance summary of the proposed methods on natural sequences is presented in Table 3. Three major conclusions can be drawn from the results in Table 3. First, HMM-based method generally outperforms MC-based one across all metrics for natural sequences. Second, first-order HMM has the best performance in terms of precision

(0.7980) and F-measure (0.7989), while second-order HMM achieves the highest recall (0.8449). Third, it follows that second-order MC and HMM have lower precision and F-measure, but higher recall than first-order MC and HMM. In particular, precision and F-measure decrease by 6.74% and 2.39%, in case of MC, and by 7.27% & 2.09%, in case of HMM, when going from first- to second-order models. In contrast, recall improves by 3%–5%, in case of both MC and HMM, when going from first- to second-order models.

Table 3: Performance of MC and HMM for predicting success of natural patient-provider communication sequences. The largest value for each performance metric is highlighted in bold.

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Markov Chain 1^{st} Order	0.7013	0.7037	0.7013	0.7003
Markov Chain 2^{nd} Order	0.5905	0.5917	0.5905	0.5890
Hidden Markov Model	0.5439	0.5778	0.5439	0.4865
LSTM RNN	0.8770	0.8812	0.8770	0.8766
LSTM RNN - TR	0.8824	0.8890	0.8824	0.8817
GRU RNN	0.8844	0.8897	0.8844	0.8840
GRU RNN - TR	**0.8858**	**0.8917**	**0.8858**	**0.8853**

Predictive performance for over-sampled PPC code sequences

Table 4 summarizes predictive performance of the proposed methods on alternating sequences. These results indicate that second-order HMM had better recall than MC, achieving 97.13% recall. Similar to the case of natural sequences, prediction method based on second-order HMM has the highest recall of 0.9713 among all models. However, different from natural sequences, second-order MC has the highest precision (0.9778) as well as the highest F-measure (0.9736). We also observed that second-order MC and HMM outperform first-order MC and HMM. Specifically, moving from first-order to second-order MC improves precision, recall and F-measure of the MI success prediction method by 0.01%, 2.72% and 1.37%, respectively. Moving from first-order to second-order HMM improves precision, recall and F-measure of the MI success prediction method by 3.32%, 4.30% and 3.70%, respectively.

Table 4: Performance of MC and HMM for predicting success of natural patient-provider communication sequences. The largest value for each performance metric is highlighted in bold.

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Markov Chain 1^{st} Order	0.7930	0.8127	0.7930	0.7896
Markov Chain 2^{nd} Order	0.7325	0.7461	0.7325	0.7288
Hidden Markov Model	0.7763	0.8062	0.7763	0.7706
LSTM RNN	0.8703	0.8782	0.8703	0.8696
LSTM RNN - TR	0.8713	0.8789	0.8713	0.8706
GRU RNN	0.8711	0.8785	0.8711	0.8705
GRU RNN - TR	**0.8722**	**0.8790**	**0.8722**	**0.8716**

Common patterns

Table 5 provides examples of typical PPC sequences that frequently appear in successful and unsuccessful motivational interviews. It can be seen that most successful patterns start with a summary of the discussion or open-ended/close-ended question. After that, if adolescents express positive change talk, the counselor immediately reflects on that to reinforce adolescent's intrinsic motivation about behavior change. On the other hand, providing information can lead to negative change talk, even in the cases when adolescents were showing positive tendency in their previous communication. This observation can be explained by adolescents quickly loosing focus when provided with general information that undermines their motivation. Analyzing such cases will allow the counselors to determine the general

information that can be provided during the interviews.

Table 5: Common patterns in successful and unsuccessful motivational interviews.

Type	Pattern
successful	328 SUM-S: Summarize → 117 LUP+: Low Uptake, positive → 313 R-CML+: Reflect, commitment language positive → [positive commitment]
successful	307 SPT: Support → 117 LUP+: Low Uptake, positive → 313 R-CML+: Reflect, commitment language positive → [positive commitment]
successful	306 CQ-EF: Closed question, Elicit Feedback → 120 HUP-O: High Uptake, other → 313 R-CML+: Reflect, commitment language positive → [positive commitment]
unsuccessful	311 R-CT+: Reflect, change talk positive → 117 LUP+: Low Uptake, positive → 302 G-INFO+: General Information, positive → [negative commitment]
unsuccessful	302 G-INFO+: General Information, positive → 117 LUP+: Low Uptake, positive → 302 G-INFO+: General Information, positive → [negative commitment]
unsuccessful	305 EA: Emphasize Autonomy → 117 LUP+: Low Uptake, positive → 302 G-INFO+: General Information, positive → [negative commitment]

Discussion

By analyzing the experimental results of different communication sequence outcome prediction schemes proposed in this paper, we arrived at the following conclusions. First, using higher-order models results in better prediction accuracy compared to lower-order models, since higher-order models utilize larger context and can better capture different nuances of patient-provider communication. On the other hand, the number of states in higher-order Markov models grows exponentially with their order. Therefore, accurate estimation of transition probabilities requires much larger training set. Using smaller datasets will result in a sparsity problem, when many transitions are either not observed in the training set at all or observed only a few times, leading to missing or potentially inaccurate probability estimates. Obtaining large training sets cannot be easily accomplished in many domains, including motivational interviewing. In this study, we found out that using second-order Markov models is a reasonable trade-off between efficiency and accuracy.

Second, the overall predictive performance of HMM is substantially better than MC for both types of sequences. In particular, HMM-based method achieves near-human accuracy for predicting the success of motivational interviews. This indicates that hidden states in HMM are able to capture the structure of discourse in motivational interviews by grouping together the codes that correspond to different cognitive states of adolescents during motivational interviews, which reflect the overall progression of the interviews. This allows to reduce the dimensionality of codes in PPC sequences and consequently improve both precision and recall of the prediction method.

Third, converting natural PPC communication sequences to alternating ones results in better performance for all configurations of the proposed method. This indicates that alternating sequences emphasize the dependencies between the pairs of patient and provider codes, which results in more accurate estimates of the corresponding conditional probabilities. Better estimates of parameters in Markov models of successful and unsuccessful interviews are propagated to the next step, where they are aggregated into predictions for the entire sequence. This allows to achieve a dramatic improvement in the prediction accuracy of the method.

Fourth, the proposed method can be used to identify the most effective communication strategies at eliciting a particular type of behavioral response. Awareness of these strategies by researchers can significantly decrease the time and effort required to develop effective interventions to address many public health conditions, such as childhood obesity, and tailor these interventions to particular patient cohorts. Awareness of these strategies by the counselors can lead to greater success rate of motivational interviews.

Conclusion

In this paper, we proposed Recurrent Neural Networks with two baseline methods Markov Chain and Hidden Markov Model for predicting the success of motivational interviews. We found out that individual patient-provider communication exchanges are highly indicative of the overall progression and future trajectory of clinical interviews and can be used to predict their overall success. Our proposed method can facilitate motivational interviewing researchers in establishing causal relationships between different communication strategies and the desired behavioral outcomes during the interviews without resource intensive manual qualitative analysis of interview transcripts. Our proposed method can also help to identify specific patterns that are common to successful and unsuccessful motivational interviews, which can also directly inform clinical practice. This work also has broad implications for qualitative public health research by providing a formal theoretically-grounded computational mechanism to facilitate the development of effective behavioral interventions.

Acknowledgments

This study was supported by an R21 grant DK108071 from the NIH. We would like to thank the student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their help with transcribing the recordings of motivational interviews.

References

[1] Leslie C, Kuang R. Fast string kernels using inexact matching for protein sequences. Journal of Machine Learning Research. 2004;5(Nov):1435–1455.

[2] Chuzhanova NA, Jones AJ, Margetts S. Feature selection for genetic sequence classification. Bioinformatics. 1998;14(2):139–143.

[3] Ye L, Keogh E. Time series shapelets: a new primitive for data mining. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2009. p. 947–956.

[4] Kudenko D, Hirsh H. Feature generation for sequence categorization. In: AAAI/IAAI; 1998. p. 733–738.

[5] Lesh N, Zaki MJ, Ogihara M. Mining features for sequence classification. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 1999. p. 342–346.

[6] Nallam K, et al. An Effective Implementation of Hierarchical Approach For Sequence Classification. IJACTA. 2016;3(2):143–146.

[7] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and knowledge discovery. 2003;7(4):349–371.

[8] Keogh EJ, Pazzani MJ. Scaling up dynamic time warping for datamining applications. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2000. p. 285–289.

[9] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257–286.

[10] Yakhnenko O, Silvescu A, Honavar V. Discriminatively trained markov model for sequence classification. In: Data Mining, Fifth IEEE International Conference on. IEEE; 2005. p. 8–pp.

[11] Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2002. p. 417–431.

[12] Wei L, Keogh E. Semi-supervised time series classification. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 748–753.

[13] Lane T, Brodley CE. Temporal sequence learning and data reduction for anomaly detection. ACM Transactions on Information and System Security (TISSEC). 1999;2(3):295–331.

[14] Drew J, Hahsler M, Moore T. Polymorphic malware detection using sequence classification methods and ensembles. EURASIP Journal on Information Security. 2017;2017(1):2.

[15] Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. Jama. 2012;307(5):483–490.

[16] General US. Surgeon Generals vision for a healthy and fit nation. Washington, DC: HHS. 2010;.

[17] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, BroganHartlieb K. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 785.

[18] Hasan M, Kotov A, Carcone AI, Dong M, Naar S, Hartlieb KB. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of biomedical informatics. 2016;62:21–31.

[19] Carcone AI, Naar-King S, Brogan K, Albrecht T, Barton E, Foster T, et al. Provider communication behaviors that predict motivation to change in black adolescents with obesity. Journal of developmental and behavioral pediatrics: JDBP. 2013;34(8):599.

[20] Eide H, Quera V, Graugaard P, Finset A. Physician–patient dialogue surrounding patients expression of concern: applying sequence analysis to RIAS. Social Science & Medicine. 2004;59(1):145–155.

[21] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357.

[22] Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms. 2011;3(1):4–21.

[23] Mutsam N, Pernkopf F. Maximum margin hidden Markov models for sequence classification. Pattern Recognition Letters. 2016;77:14–20.

[24] Eickeler S, Kosmala A, Rigoll G. Hidden markov model based continuous online gesture recognition. In: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on. vol. 2. IEEE; 1998. p. 1206–1208.

[25] Srivastava PK, Desai DK, Nandi S, Lynn AM. HMM-ModE–Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC bioinformatics. 2007;8(1):104.

[26] Won KJ, Prügel-Bennett A, Krogh A. Training HMM structure with genetic algorithm for biological sequence analysis. Bioinformatics. 2004;20(18):3613–3619.

[27] Chai W, Vercoe B. Folk music classification using hidden Markov models. In: Proceedings of International Conference on Artificial Intelligence. vol. 6. sn; 2001. .

[28] Bengio Y, Frasconi P, Simard P. The problem of learning long-term dependencies in recurrent networks. In: Neural Networks, 1993., IEEE International Conference on. IEEE; 1993. p. 1183–1188.

[29] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[30] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE; 2013. p. 6645–6649.

[31] Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259. 2014;.

[32] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014;.

[33] Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:151103677. 2015;.

[34] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference; 2016. p. 301–318.

[35] Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: Artificial Intelligence and Statistics; 2015. p. 562–570.