# Prediction of Conotoxin Type Based on Long Short-term Memory Network

**Feng Wang**
Changzhou University Huaide College
**Shan Chang**
Jiangsu University of Technology
**Dashun Wei** ( wds1149541194@163.com )
Huaide College of Changzhou University

**Research**

**Keywords:** Conotoxin, LSTM, prediction

**Posted Date:** March 30th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-273779/v1

# Abstract

**Background:** Conotoxin is a valuable peptide that targets ion channels and neuronal receptors. The toxin has been proven to be an effective drug for treating a series of diseases, but the process of identifying the type of toxin through traditional wet experiments is very complicated, low efficiency and high cost, but the method of machine learning is used to identify the cono toxin. Training in the process can effectively change this status quo.

**Methods**: A method to predict the type of spiral toxin using the sequence information of the toxin combined with the long-term short-term memory network (LSTM) method model. This method only needs to take the conotoxin peptide sequence as input, and uses the character embedding method in text processing to automatically map the sequence to the feature vector representation, and extract the features for training and prediction.

**Results**: Experimental results show that the correct index of this method on the test set reaches 0.80, and the AUC (area under the ROC curve) value reaches 0.817. For the same test set, the AUC value of the KNN algorithm is 0.641, and the AUC value of the method proposed in this paper is 0.817.

**Conclusions**: The algorithm does not require manual feature extraction and feature reconstruction steps, thereby simplifying the algorithm design, and can use the advantages of the long-term dependence of LSTM according to the characteristics of the cono toxin sequence, so that its classification can be better predicted, and the classification of the cono toxin can be better predicted. The sequence information of spirotoxin combined with the LSTM method can be better than the KNN classification algorithm.


# Background

Conus is a kind of poisonous carnivorous tropical sea and ocean soft-body animals[1]. There are more than 500 species of Conus in the world, and there are at least 50,000 active peptides in the venom of Conus. The secreted toxin (called conotoxin) is mainly used in the predation and defense behavior of animals[2]. Conotoxin is extremely toxic and can cause animals to tremble, convulse, even paralyze and die. There are more than 700 kinds of conos in the world that secrete more than 100,000 toxins. However, the current experiments have only confirmed and recorded relatively few conotoxins (about 3,000 peptides)[3]. Conotoxin has strong biological activity and novel chemical structure. It has extremely high selectivity for ligand gates or voltage-gated ion channels[4]. It can distinguish between similar ion channel types and is widely used as an ion. Pharmacological reagents in channel research. Because the insectivorous conotoxin can kill many kinds of worms[5], it has the potential to cultivate new varieties of insect-resistant crops or develop it as a peptide insecticide. Therefore, conotoxin has become a new source of new drug development and a powerful tool for pharmacology and neuroscience[6], and it ranks first in the research of animal neurotoxins. It is called "the treasure house of marine drugs", and it has received attention from all walks of life and has broad development prospects.

According to the different target sites of conotoxin[7], it can be divided into three categories: (1)Conotoxin that acts on ligand-gated ion channels. (2)Conotoxin acts on voltage-gated ion channels, which are also called voltage-sensitive channels. (3)CTX acting on other receptors[8]. There are more than 300 ion channels in living cells. Many important functions in life, such as heartbeat, sensory conduction and central nervous system response, are controlled by cell signaling through various ion channels. Ion channel dysfunction can cause a variety of diseases, such as epilepsy, arrhythmia and type II diabetes. These diseases are mainly treated with drugs that regulate the relevant ion channels[9]. Ion channels are also an important target for the treatment of viral diseases. Due to their importance to human life, ion channels have become the second most common drug development target. The following three ion channels are usually targets of toxins: potassium (K) channels,sodium (Na) channels, and calcium (Ca) channels. Based on its function and target object, conotoxin can be divided into the following three types: (i) K channel targeting type; (ii) targeting non-channel type; (iii) calcium channel targeting type[10].

Due to the explosive growth of protein sequence data[11], traditional wet experiment methods can no longer meet the needs of rapid identification of protein sequences. Yuan et al. developed a feature selection technique based on binomial distribution to predict ion channels by using radial basis function networks The type of toxin targeted. Subsequently[12], they developed a predictor (iCTX type) to improve prediction accuracy. Zhang et al. applied mixed features in the prediction problem. Wang et al. combined variance and correlation (AVC) analysis with support vector machines to reduce attribute redundancy and improve prediction accuracy and calculation speed. However, none of these methods can be used to predict the type of conotoxin defined by its target ion channel. For example, δ-toxoid-like Ac6.1 and ω-toxin-like Ai6.2 both belong to toxoid C1. However, the former targets voltage-gated sodium channels, while the latter targets voltage-gated calcium channels[13].

To solve this problem, this article proposes a method to identify the three types of conotoxins by using their sequence information alone. In this research, we propose a deep learning long-term short-term memory (LSTM) neural network model to predict the classification of cono toxins[14], and use word embedding technology to represent the conotoxin sequence as a vector, which is because the protein sequence can be seen Into a natural language. Effective features are extracted from the conotoxin sequence in order to further evaluate the performance of the model. The target model is compared with the existing machine learning model SVM[15]. The experimental results show that the method has good prediction performance and is suitable for classification and prediction of conotoxin. The workflow is shown in Fig. 1.

In this paper, word embedding technology and LSTM are combined to construct a model for anticancer peptide prediction, so as to take advantage of LSTM's advantages in sequence modeling and long-term memory and word embedding in sequence representation.

# Results

# Experiment1:Model parameter optimization

Due to the small number of training sets, this paper adopts the cross-checking method to conduct experiments. In order to conduct effective verification, accuracy and ROC are used as measurement indicators[22].

Because this paper predicts which ion channel the conotoxin belongs to is a three classification problem, that is, according to the sequence of the conotoxin to determine whether it belongs to a potassium ion channel, a sodium ion channel or a calcium ion channel, the activation function is softmax when compiling the model[23].

First, determine the appropriate word vector embedding dimension. According to the characteristics of the collected conotoxin sequence data, the dimension of the word embedding vector space is selected as 60 and 90 for comparative analysis. The ROC curve corresponding to the model during verification is shown in Fig. 3.

# Experiment2:Distribution of accuracy and loss function on training set and test set

# Experiment3: Algorithm performance comparison

KNN is one of the most commonly used classification algorithms. It has a good predictive effect and is not sensitive to outliers. Considering that there may be some erroneous data in the conotoxin data set collected in this paper, the KNN algorithm is used as a comparison algorithm. The ROC curves of the two methods on the test set are shown in Fig. 5. It can be seen from Fig. 5 that the area under the LSTM curve is larger than the area under the KNN curve, indicating that the accuracy of LSTM is higher than that of KNN, which proves that the method based on LSTM is superior to the KNN algorithm Dealing with the problem of conotoxin data classification.

# Discussion

When the embedding dimension is 90, the obtained area under the ROC curve, that is, the AUC value, is the largest. At this time, the prediction performance is the best. Therefore, the vector dimension of the word embedding space is set to 90,Fig. 3. When the training parameter epoch is set to 10 times, the accuracy and loss function curves of the model on the training set and independent test set are shown in Fig. 4. As can be seen from Fig. 4, whether it is on the training set or the test set, the accuracy and loss value curves are close to each other, indicating that there is no over-fitting phenomenon, which shows that the model has a good generalization ability,Fig. 4. Combined with Fig. 4(a), it can be concluded that considering the imbalance of classification data, the method proposed in this paper has certain reference value for both accuracy and ROC, which further shows that the method is in the treatment of conotoxin. The superiority of the three classifications,Fig. 5.

## Conclusion

According to the characteristics of the conotoxin sequence, this paper uses the LSTM algorithm based on the word embedding method to classify and predict the conotoxin. The algorithm does not require manual feature extraction and feature reconstruction steps, which simplifies the algorithm design, and can use the advantages of the long-term dependence of LSTM according to the characteristics of the conotoxin sequence to provide a better prediction for the classification of the conotoxin Performance and experimental results show that the proposed algorithm can effectively predict the conotoxin in three categories.

## Methods

# Data set

The conotoxin sequence and its function used in this experiment are collected from UniProt[16]. In order to improve the quality of the data, when collecting data, we first limit the function of the conotoxin to support potassium, calcium and sodium channels[17]. There are no conotoxins clearly marked on UniProt. Except for a few of the conotoxins we have identified, all the others are discarded. In the end, we obtained 192 conotoxins, of which 74 calcium ion channel targeting types, 84 sodium ion channel targeting types, and 34 potassium ion channel targeting types. The training set consists of 60 calcium ion channel conotoxins, 67 sodium ion channel conotoxins and 25 potassium ion channel conotoxins. The test set consists of 14 calcium ion channel conotoxins and 17 sodium ion channel conotoxins. Toxin and 9 kinds of potassium ion channel conotoxin. The details of the data set are shown in Table 1.

Table 1
Summary of the data set

| Data set | Ca ion channel | Na ion channel | K ion channel |
|---|---|---|---|
| Training set | 60 | 67 | 25 |
| Test set | 14 | 17 | 9 |

# Sequence characterization

This algorithm does not need to manually determine the physical and chemical properties of amino acids through wet experiments. It only uses the conotoxin character sequence as input data, and uses the word embedding training method to divide the conotoxin sequence into individual characters; because the length of the conotoxin sequence is not fixed, So we set a fixed maximum length according to the data set, and encode the conotoxin sequence with a fixed length. When the length of the encoded sequence is less than the maximum fixed length, fill it with 0 at the end, so that each character corresponds to An integer; then the word embedding training is carried out through the neural network, and 20 amino acid letters are mapped to the word embedding vector space, so that each character corresponds to a vector representation. The above steps can be automatically completed by the Tokenizer API provided by Keras.

Each conotoxin sequence can be coded as an M × N matrix, where M is the set sequence length and N is the set embedding space vector dimension.

# LSTM three-class prediction model

LSTM is a recurrent neural network with a special structure, which is an effective technology to solve the problem of long sequence dependence[18]. It is composed of a group of unit modules with memory function. Each unit module is composed of input gate, forget gate and output gate to realize the input, filtering and output of information. These gated operations enable LSTM to automatically extract and learn long-range correlation information useful for the overall classification task in the sequence, and the prediction of conotoxin classification based on sequence information is just in line with the characteristics of this type of sequence classification problem, so LSTM is suitable for Classification of Conotoxin.

This article classifies three ion channel-targeted conotoxins of potassium ion, calcium ion and sodium ion. Therefore, the activation function should not use the sigmoid function[19], but the softmax activation function. This is because the effect of sigmoid in dealing with two classification problems Not bad, but the softmax function works better when dealing with multi-classification problems.

$$S_i = \frac{e^{z_i}}{\sum_k e^{z_k}} \qquad (1.1)$$

The overall process of the classification prediction algorithm proposed in this paper is shown in Fig. 2. First, the amino acid characters appearing in the conotoxin sequence are automatically mapped to the embedding vector space after neural network training, so that each amino acid character corresponds to a vector representation; then each conotoxin sequence is represented as a corresponding matrix; finally, The matrix is used as the input of the LSTM model for training and learning.

# Evaluation method and evaluation index

Cross-validation and independent test data sets are used to verify the performance of the algorithm in this paper. Cross-validation divides the training set data into five sub-sets[20]. Each time one subset is used as the test set for verification, and the remaining four combinations are combined as the training set. This process is repeated 5 times until each subset is considered as a test set at least once. At the same time, this paper also uses an independent test data set to verify the performance of the algorithm. The evaluation indicators of the algorithm include: 1) True Positive Rate (TPR); 2) False Positive Rate (FPR); 3) Correct Index; 4) ROC[21] curve and the area value AUC under it. The calculation formula for each indicator is as follows:

$$TPR = \frac{TP}{TP+FN} \times 100\% \quad (1.2)$$

$$FPR = \frac{FP}{TN+FP} \times 100\% \quad (1.3)$$

$$Correct\ index = (TPR + 1 - FPR) \times \frac{1}{2} \quad (1.4)$$

In the formula: TP refers to the number of positive samples predicted to be positive; FP refers to the number of negative samples predicted to be positive; TN refers to the number of negative samples predicted to be negative; FN refers to the number of positive samples predicted to be negative.

## Declarations

# Compliance with Ethical Standards

**Research involving human participants and/or animals:** This article does not contain any studies with human participants performed by any of the authors.**Funding:** There is no funding for this study
**Conflict of Interest:** The authors declare that they have no conflict of interest.

**Acknowledgements:**Not applicable.

**Funding:** There is no funding for this study.

**Authors' contributions:** Feng Wanga,and Dashun Weia conceived and wrote the main sections of the review. Shan Chang contributed to some sections and criti-cally reviewed the manuscript. All authors read and approved the final manuscript.
**Availability of data and materials:**Not applicable.

**Ethics approval and consent to participate:**Not applicable.

**Consent for publication:**Not applicable.

**Competing interests:**The authors declare no competing interests.

# References

1. Conus is a. kind of poisonous carnivorous tropical sea and ocean soft-body animals.
2. Julien G, David W, Annette N, et al. Synthesis, Structure and Biological Activity of CIA and CIB, Two α-Conotoxins from the Predation-Evoked Venom of Conus catus[J]. Toxins. 2018;10(6):222.
3. Adams DJ, Alewood PF, Craik DJ, et al. Conotoxins and their potential pharmaceutical applications[J]. Drug Dev Res. 2015;46(3–4):219–34.

4. Gielen M, Pierre-Jean, Corringer. The dual-gate model for pentameric ligand-gated ion channels activation and desensitization[J]. The Journal of Physiology, 2018, 596(10).
5. Manhães MA, Dias MM, Lima ALC. Feeding resource partitioning between two understorey insectivorous birds in a fragment of Neotropical cloud forest[J]. Brazilian journal of biology = Revista brasleira de biologia. 2015;75(4 suppl 1):176.
6. Rajabi H, Zolgharnein H, Ronagh MT, et al. Conus coronatus and Conus frigidus Venom: A New Source of Conopeptides with Analgesic Activity[J]. Avicenna Journal of Medical Biotechnology, 2020, 12(3).
7. Wu RJ, Wang L, Xiang H. The Structural Features of α-Conotoxin Specifically Target Different Isoforms of Nicotinic Acetylcholine Receptors[J]. Current Topics in Medicinal Chemistry, 2016.
8. Zhang J, Chen X, Xue Y, et al. Beyond voltage-gated ion channels: Voltage-operated membrane proteins and cellular processes[J]. Journal of Cellular Physiology, 2018.
9. Domínguez-Mozo MI, Toledano-Martínez, et al. JC virus reactivation in patients with autoimmune rheumatic diseases treated with rituximab.[J]. Scandinavian journal of rheumatology, 2016.
10. Rusconi F, Ceriotti P, Miragoli M, et al. Therapeutic modulation of cardiac function by selective peptidomimetic-mediated targeting of the l-type calcium channel machinery[J]. Vascul Pharmacol. 2015;75:55–6.
11. Mason R. Application of Cathodoluminescence Imaging to the Study of Sedimentary Rocks[J]. J Geol. 2006;115(6):710–0.
12. Li W, Xing Z, Guanghui C, et al. Data hierarchical fusion release mechanism based on differential privacy protection [J]. Minicomputer system. 2019;40:(10).
13. Semwal VB, Singha J, Sharma PK, et al. An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification[J]. Multimedia Tools Applications. 2016;76(22):24457–75.
14. Son H, Paul A, Jeon G. Country Information Based on Long-Term Short-Term Memory (LSTM)[J]. International Journal of Engineering Technology. 2018;7(4.44):47.
15. Guo H, Wang W.. An active learning-based SVM multi-class classification model[J]. Pattern Recognition, 2015, 48(5).
16. Consortium UP. UniProt: a hub for protein information[J]. Nuclc Acids Research(D1):204–12.
17. Murenzi E. Evaluation of microtransplantation of rat brain neurolemma into Xenopus laevis oocytes as a technique to study the effect of neurotoxicants on endogenous voltage-sensitive ion channels[J]. NeuroToxicology. 2017;60:260–73.
18. Agarap AF. A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data[J]. 2017.
19. Tsai CH, Chih YT, Wong WH, et al. A Hardware-Efficient Sigmoid Function With Adjustable Precision for a Neural Network System[J]. IEEE Transactions on Circuits Systems II Express Briefs. 2017;62(11):1073–7.

20. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning[J]. Springer Open Choice, 2018, 2(3).
21. Hoyer A, Hirt S, Kuss O.. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data[J]. Research Synthesis Methods, 2018, 9(1).
22. Dang TK, Pham DMC, Ho DD.. On verifying the authenticity of e-commercial crawling data by a semi-crosschecking method[J]. International Journal of Web Information Systems, 2019(2).
23. Gimpel K, Smith NA.. Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions[J]. 2010.

# Figures

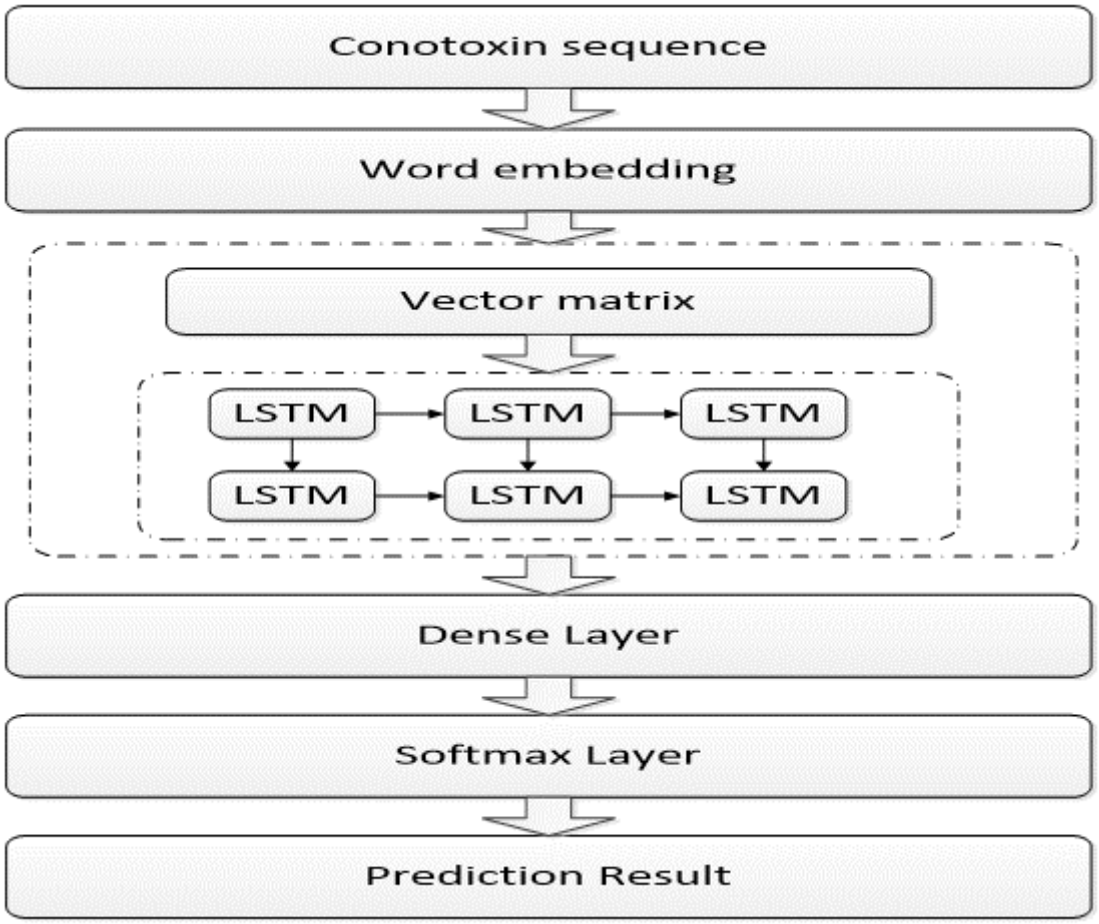



**Figure 1**

The Flowchat of Proposed method

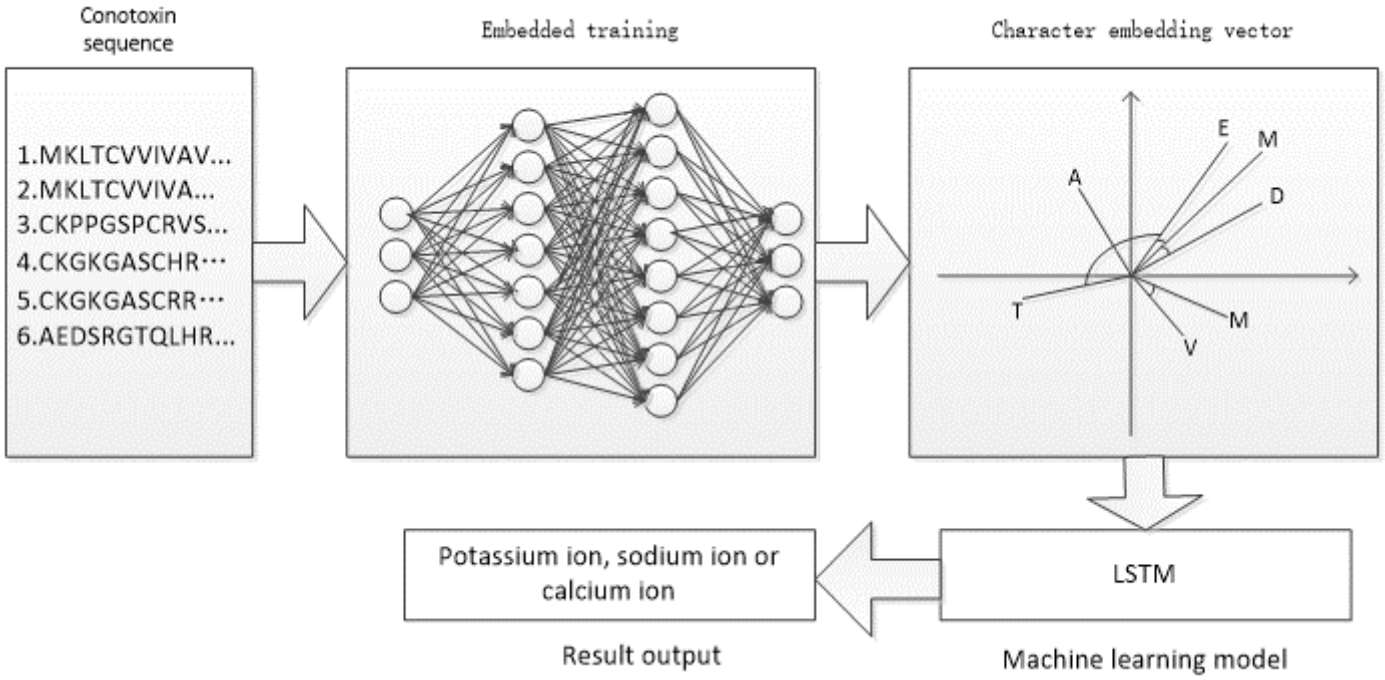



**Figure 2**

Algorithm flow

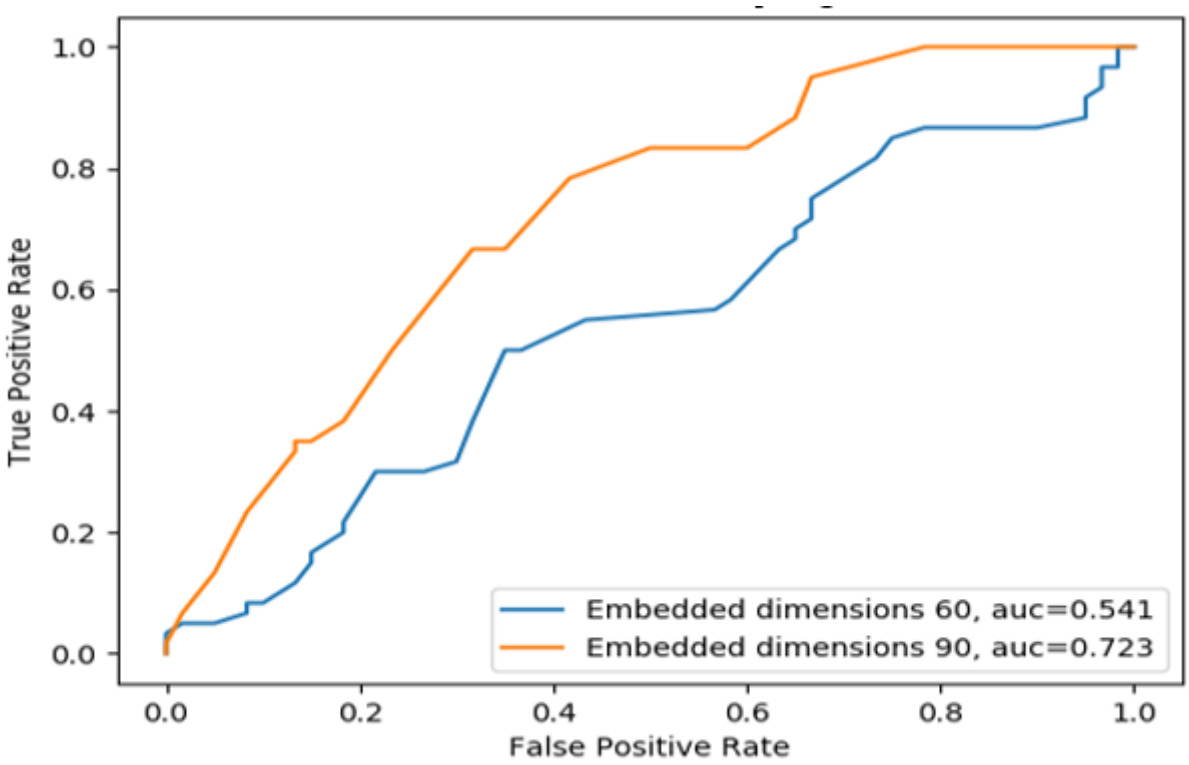
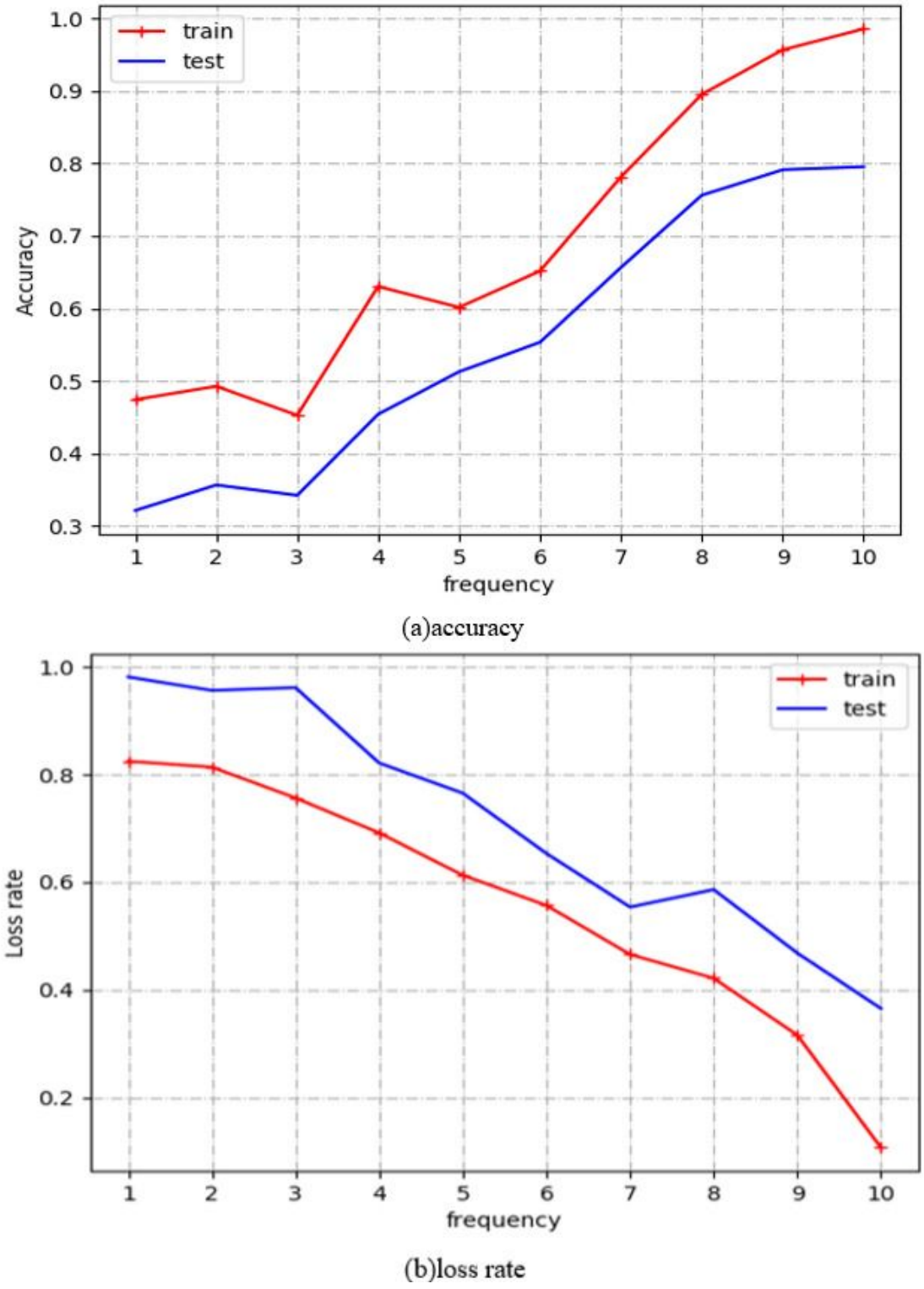



**Figure 3**

ROC curve corresponding to different word embedding dimension




## Figure 4

Model accuracy curve and loss function curve

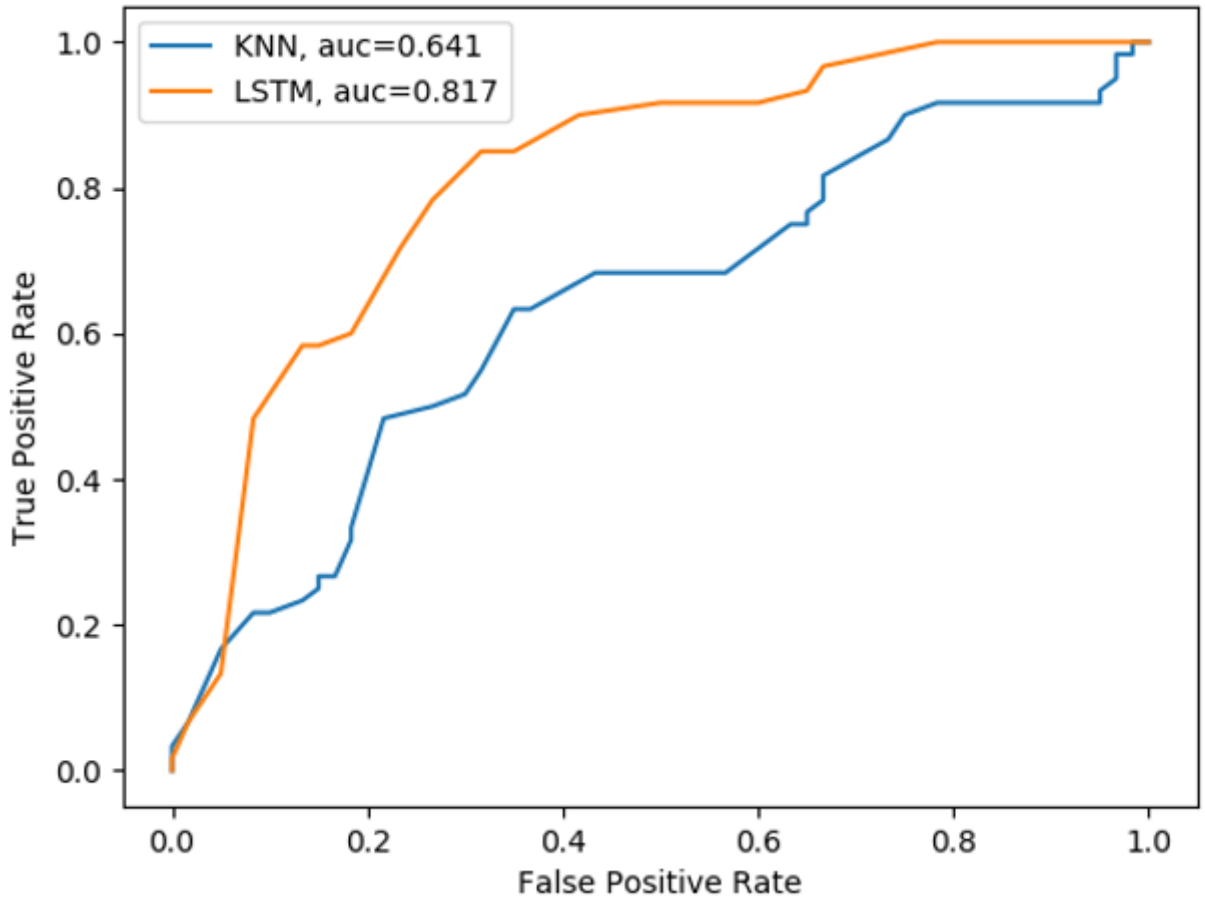



**Figure 5**

ROC curve of LSTM and KNN on independent test set