**Proceedings Paper:**
Aftab, Haris orcid.org/0000-0001-7981-1743, Gautam, Vibhu, Hawkins, Richard David orcid.org/0000-0001-7347-3413 et al. (2 more authors) (2022) Robust Intent Classification using Bayesian LSTM for Clinical Conversational Agents (CAs). In: MobiHealth 2021:Wireless Mobile Communication and Healthcare. 10th EAI International Conference on Wireless Mobile Communication and Healthcare, 13-14 Nov 2021 Springer , CHN , pp. 106-118.

https://doi.org/10.1007/978-3-031-06368-8_8

# Robust Intent Classification using Bayesian LSTM for Clinical Conversational Agents (CAs)

Haris Aftab [0000-0001-7981-1743], Vibhu Gautam[0000-0001-9040-9843], Richard Hawkins[0000-0001-7347-3413], Rob Alexander[0000-0003-3818-0310], and Ibrahim Habli [0000-0003-2736-8238]

Department of Computer Science, University of York, York, YO10 5GH, United Kingdom
haris.aftab@york.ac.uk

**Abstract.** Conversational Agents (CAs) are software programs that replicate human conversations using machine learning (ML) and natural language processing (NLP). CAs are currently being utilised for diverse clinical applications such as symptom checking, health monitoring, medical triage and diagnosis. Intent classification (IC) is an essential task of understanding user utterance in CAs which makes use of modern deep learning (DL) methods. Because of the inherent model uncertainty associated with those methods, accuracy alone cannot be relied upon in clinical applications where certain errors may compromise patient safety. In this work, we employ Bayesian Long Short-Term Memory Networks (LSTMs) to calculate model uncertainty for IC, with a specific emphasis on symptom checker CAs. This method provides a certainty measure with IC prediction that can be utilised in assuring safe response from CAs. We evaluated our method on in-distribution (ID) and out-of-distribution (OOD) data and found mean uncertainty to be much higher for OOD data. These findings suggest that our method is robust to OOD utterances and can detect non-understanding errors in CAs.

**Keywords:** Conversational Agents (CAs), Machine Learning, Model Uncertainty, Out-of-Distribution (OOD), Healthcare, Patient Safety.

## 1 Introduction

Conversational Agents (CAs) such as Google Home and Amazon Alexa are interactive conversational systems that use Machine Learning (ML) to respond to the user in natural language via voice or text [1]. They can be categorised into two types: task-oriented CAs [2] and chatbots [3]. In healthcare studies, task-oriented CAs are often utilised as they are focused on achieving a task such as booking a consultation or finding a hospital. Chatbots are systems designed for open-ended conversations and mimic unstructured conversations or chats. Common applications of CAs in healthcare include symptom checking [4], chronic disease management [5], health monitoring and medication adherence [6].

CAs employ a pipeline architecture [7] as shown in Figure 1. The fundamental components in this architecture are Natural Language Understanding (NLU) and Dialog Manager (DM) which enable their understanding and decision making. The user then

receives the response proposed by DM via the Natural Language Generation (NLG) module. In this pipeline architecture, the NLU maps user utterances to intents and slots and has a significant impact on downstream processing. NLU errors may lead to erroneous decision making [8], which can be costly in healthcare because of the risk to human life and ethical issues [9]. Specifically, the NLU in CAs is concerned with the IC and slot-filling (SF) [7]. IC predicts a user's intent from a given utterance, and it is a classification problem of identifying the correct intent label. SF in NLU extracts additional information needed to accomplish the user's task. For example, a user asking a CA *"show me nearby hospitals"* could have *'show_hospital'* as intent and the current user location as the slot value.

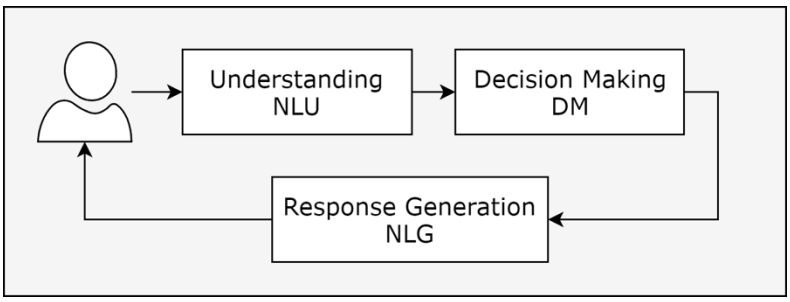


Figure 1: Conversational Agent (CA) Architecture

DL have allowed significant performance enhancements in computer vision and Natural Language Processing (NLP) tasks and their variants such as Recurrent Neural Networks (RNNs) [10, 11] and Long Short-Term Memory Networks (LSTMs) [12] are commonly used for IC in CAs. These networks are able to attain higher accuracy on text classification tasks as they are better suited to model time series data.

Existing state-of-the-art Deep learning (DL) methods are prone to data and model uncertainties [13]. Model uncertainty, also known as epistemic uncertainty, occurs because of the reliance of the model on training data for their prediction. This uncertainty can be reduced by providing enough training data. Estimating model uncertainty is extremely crucial also because of the difficulty to obtain high-quality datasets in healthcare [14]. In addition, it is almost impossible to provide complete data as DL models will always reflect an imperfect representation of the real world [15].

In general, for classification problems, the softmax function is utilised by DL models at the output, resulting in a probability distribution over class labels. The label with the highest probability is then chosen as the prediction. The softmax function calculates relative probabilities between classes but does not provide a measure of the model's uncertainty [16]. The probabilistic nature of softmax output is one of the reasons this score cannot be used as a confidence measure of the model in its prediction. DL models on unseen data tend to make predictions with the high softmax values and thus it is undesirable to use them in safety-critical systems [17].

CAs are vulnerable to failures in understanding user utterance, and non-understanding errors are one of those failures [18]. Non-understanding errors arise when the system is unable to understand user input due to the system's inability to support the requested feature or poorly formatted input. For example, a user asking a COVID symptom checking CA about diabetes symptoms would result in a non-understanding error. Similarly, any unknown or incorrect input would also cause a non-understanding error. A common source of non-understanding errors is out-of-distribution (OOD) data [19]. Non-understanding errors usually result in poor user experience and may not be desirable to have them in safety-critical applications. As described earlier, the way DL models make predictions and are inherently uncertain, the need to detect non-understanding errors is significant in CAs that utilise DL methods.

Bayesian modelling techniques provide a probabilistic representation of model uncertainty but these usually are computationally expensive [16]. It is however possible to interpret DL methods as Bayesian models without modifying the model to reduce this computational complexity [20]. DL methods suffer from overfitting with limited training examples and dropouts are utilised during training time to prevent it. Additionally, these dropouts can be used at test time to generate random predictions which are sampled out to interpret in a probabilistic manner. This technique is known as Monte-Carlo (MC) dropout [16]. In this work, we apply Bayesian method to model LSTM for IC which enables us to quantify model uncertainty, thus enhancing confidence in model's decisions during IC.

The key contributions of this paper are:

1. We utilise Bayesian LSTM with MC dropout for computing uncertainty in IC for CAs.
2. A symptom checking prototype CA is designed to demonstrate the importance of robust IC in CAs and how our method can be utilised for assuring safe response.
3. We evaluate our approach using an OOD evaluation dataset and compare the results to ID data.

## 2 Related Work

IC methods in CAs range from rule-based to ML approaches, but the state-of-art in IC use DL methods which include RNNs and LSTMs [10, 21]. Westhuizen et al. [22] show the utility of Bayesian LSTMs on medical time series data using MC dropout and concluded their performance enhancements over standard LSTMs. They utilised MC dropout for 100 Bayesian LSTM samples and found that using it during test time enhanced performance on all datasets and provided the added benefit of having a confidence measure alongside the predicted class. Dusenberry et al. [9] investigated several strategies to analyse model uncertainty for electronic health records. In comparison to ensemble RNNs, Bayesian RNNs performed better while only requiring training a single model. These authors concluded that Bayesian RNNs are more efficient, making them better suited for use in medical domain.

Other studies in healthcare involving deep neural networks (DNNs) have employed MC dropout to approximate uncertainty for classification tasks [23, 24]. These, however, make use of image data to estimate uncertainty. This method is also used in other safety-critical domains such as autonomous vehicles (AV), to estimate uncertainty for the AV to make safe decisions such as decelerate to speed limit or brake to stop driving [25].

The use of Bayesian approach, in addition to providing confidence in the decision of the model, enables us to detect non-understanding errors in CAs. As mentioned already, OOD data is one of the sources of these errors. It is critical to correctly identify OOD data in NLU to avoid DM taking an incorrect action [26] which could be catastrophic. Common approaches used for OOD detection rely on a threshold measure, which is subsequently utilised to compute a detection score using various methods. Bayesian models [27], and classifier ensembles [28] are two of these approaches. However, these approaches are computationally expensive, which limits their utility in industrial settings.

Another method for determining OOD detection is to use the highest softmax value as the detection score. However, as recent research has demonstrated [16], the softmax value is not a credible indication of the model's confidence. Other approaches rely on OOD labels with training examples [29], which is not viable since we cannot estimate how many OOD samples are necessary for training a model. A few studies [30, 31] have relied on OOD data creation to boost detection scores. This necessitates the creation of OOD samples for detection and reliance on tagged instances, which is an additional step in OOD detection process.

In [22], MC dropout for classification was utilised using medical data for image and speech datasets. Unlike the work in [22], we employed text data for our classification of medical time series data and analysed the impact of misclassification on patient safety by presenting a use case of symptom checking CA. In addition, we validate our method on an evaluation dataset designed for OOD data [26] which is also used in other studies [31]. We perform a comparison of results of uncertainty estimation between ID and OOD data which is discussed in detail in the Results section.

## 3 Methods

We employ Bayesian LSTM as part of our RNN architecture for the IC model of NLU. MC dropout [16], which is used at test time is then utilised to evaluate model uncertainty for IC. We designed a use case and implemented a prototype CA that performs symptom checking on medical data. In this use case, we are concerned with how uncertainty estimation in IC in CAs can aid in assuring safe response.

### 3.1 Bayesian LSTM

Bayesian implementation of LSTM allows us to estimate model uncertainty, which indicates our imperfect understanding of the model's underlying parameters. Dropout at

test time allows us to approximate the variational posterior distribution of model parameters (weights and biases). Using random dropout, we can sample different model parameters of this posterior distribution. By introducing a distribution over all model parameters, different functions can be induced. Through the realisation of distinct model parameter values selected from the posterior distribution, these functions lead to varied outcomes. The softmax predictions from each of these sampled parameters are averaged for new data. This allows us to have increased confidence in the softmax prediction. The softmax class prediction is then used to estimate model uncertainty in the form of Shannon entropy [31].

Table 1 shows the architecture of the Bayesian LSTM we utilise for our IC model. We implemented a Bayesian LSTM layer referred to as 'MCLSTM', which allows us to employ the same dropout mask during test time at each time step of recurrent layers of LSTM [20]. A dropout rate of 70% was utilised to estimate model uncertainty. The hyperparameter, dropout, at this percentage produced the best model accuracy and robust model uncertainty. We apply MC dropout after the dense layer allowing us to capture the model uncertainty for the dense layers as well.

Table 1: Recurrent Neural Network Architecture

| Layer | Output Shape | Parameters |
|---|---|---|
| Input Layer | (None, 30) | 0 |
| Embedding | (None, 30, 50) | 5000000 |
| MCLSTM | (None, 64) | 29440 |
| Dense Layer | (None, 256) | 16640 |
| Activation | (None, 256) | 0 |
| Dropout | (None, 256) | 0 |
| Dense Layer | (None, 25) | 6425 |
| Activation | (None, 25) | 0 |

### 3.2 Symptom Checker Use Case

We present a symptom checking CA prototype to highlight the impact of incorrect IC on patient safety and how our method can aid in providing a safe response when the model is uncertain about its prediction. As an example, during the current COVID-19 pandemic, many web and mobile-based applications were developed for the general public to check if they have COVID symptoms [32]. The reliability of the decisions made by these diagnostic systems can not solely rely on their accuracy [9] and this also holds for clinicians making their decisions [33]. From the clinical safety perspective, a calibration of confidence and accuracy is important.

The architecture of our prototype CA is shown in Figure 2. The input text utterance is provided by the user, which is handled by the NLU and IC is performed using Bayesian LSTM. In the case where the NLU is not certain about the prediction, a safe strategy (asking the user to rephrase or connecting the user to a human clinician) can be utilised before the NLU result is passed to the DM.

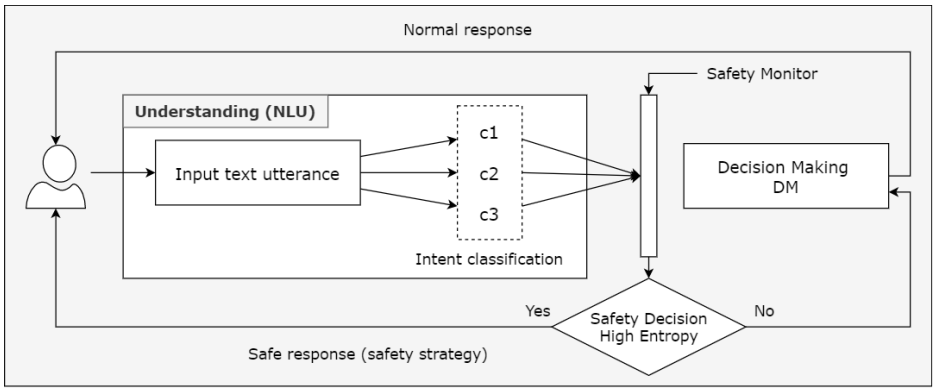


Figure 2: Symptom checker use case CA architecture diagram

We utilise an open-source dataset [34] to train our Bayesian LSTM model for understanding. The dataset contains 6661 text utterances of common medical symptoms like “knee pain”, or “headache”. The dataset contains 25 distinct intents which are evenly distributed across the dataset as shown in Figure 3. We pre-process the dataset by performing case normalization and removing punctuations and white spaces. After the pre-processing step, the utterances are padded to be of equal length. To use the data, we then transform the text utterances to numerical data using one-hot encoding scheme. We use an 85:15 ratio to split the dataset into training and testing, which turns our training size to 5661 and the test size to 1000 utterances.

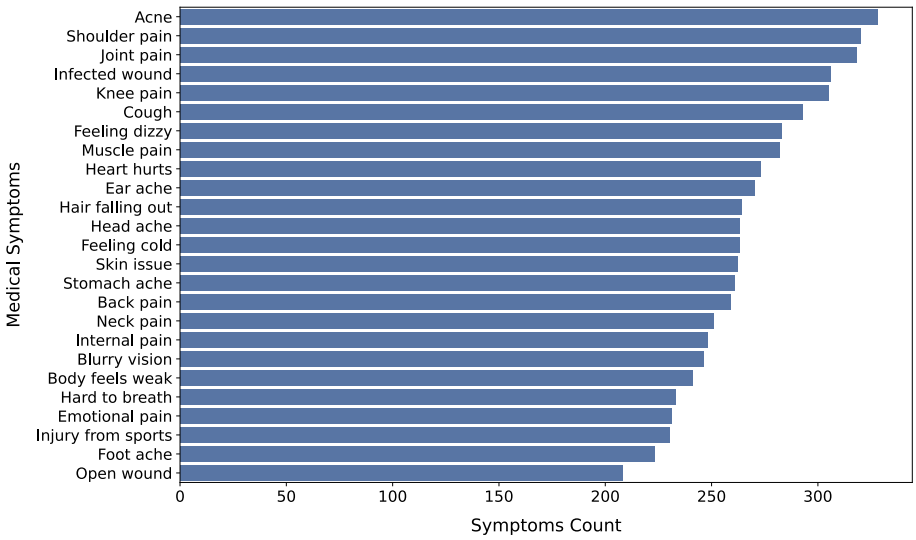


Figure 3: The distribution of medical symptoms in the dataset

## 4 Results

Our model utilising Bayesian LSTM achieved an accuracy of 99.4% on the test dataset. Figure 4 shows the confusion matrix which reflects the model's high accuracy. The y-axis lists the actual symptoms, and the x-axis lists the predicted symptoms by the model. Due to the higher accuracy, there are very few misclassifications by the model.

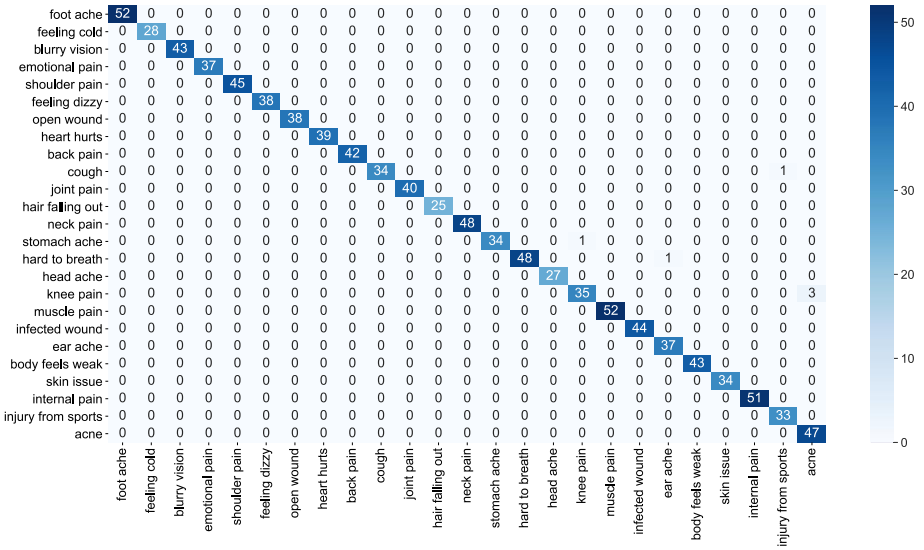


Figure 4: Confusion matrix of symptoms classification

Table 2 summarises the average findings for each of the medical symptoms (intents) in the dataset by the following evaluation metrics: precision, recall, and F1-score. The number of samples for each intent is represented by the "Support" column, which indicates that there is no class imbalance in the test set. Because of their increased accuracy, these evaluation metrics appear to indicate near-perfect scores for each of the intents. The precision and recall usually do not provide a good measure of the quality of the model as they can be high because of class imbalance. The F1-score provides a weighted average of both the precision and recall and in our experiment, it also achieves a near 100% score for most of the intents which is an indication of good model performance. The average metrics (macro and weighted average) scores indicate that there is very little class imbalance which validates the high accuracy on the test set.

We sample the softmax value for the same input 100 times to calculate the uncertainty. This yields the output posterior distribution for softmax values, which is then averaged, and the entropy for all outputs is calculated. A higher entropy value reflects high uncertainty which indicates the possibility of the input from OOD data [31]. Table 3 lists the ID utterances randomly selected from the test dataset, predictions, and their entropy calculations. The model correctly predicts all the utterances which is due to the higher model accuracy and ID nature of utterances.

Table 2: Average evaluation metrics for medical symptoms

| Medical Symptoms | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| acne | 1.000 | 1.000 | 1.000 | 52 |
| back pain | 1.000 | 1.000 | 1.000 | 28 |
| blurry vision | 1.000 | 1.000 | 1.000 | 43 |
| body feels weak | 1.000 | 1.000 | 1.000 | 37 |
| cough | 1.000 | 1.000 | 1.000 | 45 |
| ear ache | 1.000 | 1.000 | 1.000 | 38 |
| emotional pain | 1.000 | 1.000 | 1.000 | 38 |
| feeling cold | 1.000 | 1.000 | 1.000 | 39 |
| feeling dizzy | 1.000 | 1.000 | 1.000 | 42 |
| foot ache | 0.971 | 1.000 | 0.986 | 34 |
| hair falling out | 1.000 | 1.000 | 1.000 | 40 |
| hard to breath | 1.000 | 1.000 | 1.000 | 25 |
| head ache | 1.000 | 1.000 | 1.000 | 48 |
| heart hurts | 0.971 | 1.000 | 0.986 | 34 |
| infected wound | 0.980 | 1.000 | 0.990 | 48 |
| injury from sports | 1.000 | 1.000 | 1.000 | 27 |
| internal pain | 0.921 | 0.972 | 0.946 | 36 |
| joint pain | 1.000 | 1.000 | 1.000 | 52 |
| knee pain | 1.000 | 1.000 | 1.000 | 44 |
| muscle pain | 1.000 | 0.974 | 0.987 | 38 |
| neck pain | 1.000 | 1.000 | 1.000 | 43 |
| open wound | 1.000 | 1.000 | 1.000 | 34 |
| shoulder pain | 1.000 | 1.000 | 1.000 | 51 |
| skin issue | 1.000 | 0.971 | 0.985 | 34 |
| stomach ache | 1.000 | 0.940 | 0.969 | 50 |
| accuracy | 0.994 | 0.994 | 0.994 | 0.994 |
| macro avg | 0.994 | 0.994 | 0.994 | 1000 |
| weighted avg | 0.994 | 0.994 | 0.994 | 1000 |

Table 3: Uncertainty estimation for in-distribution (ID) utterances

| Test Utterance (ID) | Prediction | Entropy |
|---|---|---|
| my head is so heavy cant think normally | head ache | 0.029 |
| i feel a burning sensation in my shoulder muscle | muscle pain | 0.055 |
| i can hardly breathe | hard to breath | 0.071 |
| i have internal pain whenever i come down with a cold | internal pain | 0.327 |
| when i'm awake in the morning i feel strange and have vertigo | feeling dizzy | 0.507 |

Table 4: Uncertainty estimation for out of distribution (OOD) utterances

| Test Utterance (OOD) | Prediction | Entropy |
|---|---|---|
| am i connected to wifi | feeling cold | 1.057 |
| how much time do i have left on my 0 apr | shoulder pain | 1.110 |
| what casino game has the best odds | injury from sports | 1.862 |
| please alert me when my iphone battery falls below 30 | neck pain | 2.134 |
| what is the warranty on my microwave | skin issue | 2.302 |

Table 4 shows the five random utterances from OOD dataset [26] with model prediction and entropy calculations. This dataset contains 1000 utterances for evaluation purpose and differs from the dataset on which the IC model is trained. As shown in Figure 5, the mean entropy for this OOD dataset for an identical number of samples is 2.025 which is substantially higher than the mean entropy of 0.098 for ID utterances. This demonstrates that our method can be utilised to detect non-understanding errors as well as to help assure the safety of CA response in the wake of uncertainty from DL models.

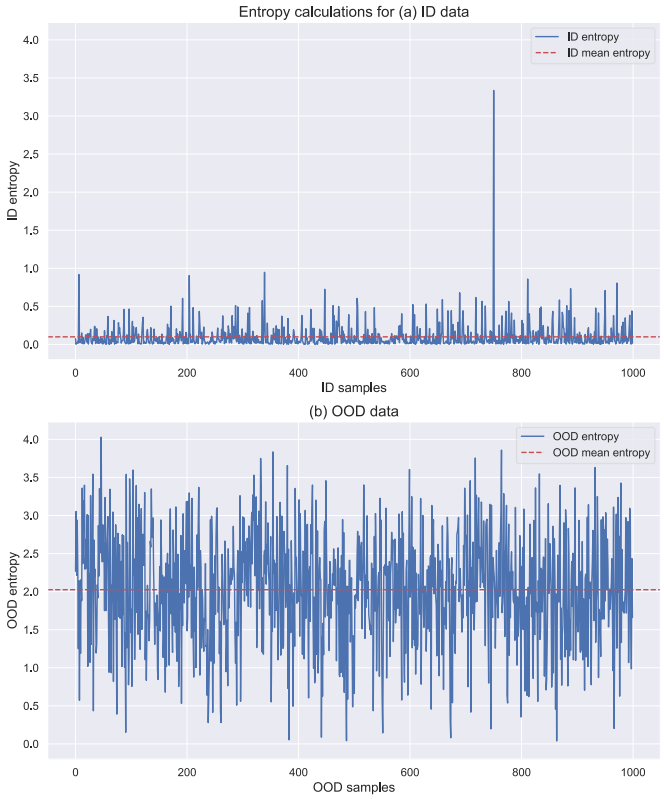


Figure 5: Entropy calculations for ID (top) and OOD (bottom) data

## 5 Discussion

Our model based on Bayesian LSTM yielded high accuracy of 99.4% on the test dataset. The training dataset examples contained low class imbalance and we applied dropout during training to improve model performance. The use of MC dropout at test time enabled us to sample multiple outputs and we calculated entropy by averaging out 100 samples from this distribution. It is worth noting that the classifier in this case even having near 100% accuracy cannot be trusted from their prediction alone which we discussed earlier. As seen in Table 4, for all OOD utterances the prediction was incorrect with high uncertainty. The average model entropy for ID data (test dataset) was much lower than the average entropy for OOD data with the same number (1000) of examples. It is yet to be seen if this pattern continues for a very large number of OOD data.

The state-of-art in CAs rely on DL methods [10] which are prone to uncertainties in their decisions [35]. In healthcare, instead of making wrong predictions, these models should be able to say "sorry, I don't know" when they are uncertain. From our findings on OOD of relatively small size (1000 samples), the entropy measure can be utilised to know when a model is uncertain in its decision. We present a use case of symptom checking where this method during IC can be useful for providing a safe response. A safety monitor such as one discussed in [36] may be deployed after NLU output which can filter high uncertainty inputs to avoid any incorrect actions by the DM. Alternatively, as mentioned in [37], a user may be asked to provide a rephrase input. In case of high uncertainty, another approach of handing over the control to a human clinician may also be used [38].

## 6 Conclusion and Future Work

In this paper, we presented a robust mechanism for IC in clinical CAs by measuring model uncertainty using Bayesian LSTMs. A symptom checking prototype CA was implemented to illustrate the benefit of certainty measure alongside prediction. This method shows that non-understanding errors in CAs can be avoided and a safety strategy (safety monitor in CA architecture, or human involvement) can be utilised to prevent unsafe responses. We evaluated our approach on a dataset of 1000 samples and the results were promising. However, further research may be required to estimate the minimum data required for this method. Additionally, data uncertainty [35] which occurs due to noise in the data may require to be calculated for the assurance of safe response in CAs.

### Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 812.788 (MSCAETN SAS). This publication reflects only the authors' view, exempting the European Union from any liability. Project website: http://etn-sas.eu/.

# 7 References


1. Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y.S., Coiera, E.: Conversational agents in healthcare: A systematic review. J. Am. Med. Informatics Assoc. 25, 1248–1258 (2018).
2. Gao, J., Galley, M., Li, L.: Neural approaches to conversational ai. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1371–1374 (2018).
3. Harms, J.-G., Kucherbaev, P., Bozzon, A., Houben, G.-J.: Approaches for dialog management in conversational agents. IEEE Internet Comput. 23, 13–22 (2018).
4. Razzaki, S., Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., Taliercio, M., Butt, M., Majeed, A., DoRosario, A., Mahoney, M., Johri, S.: A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. 1–15 (2018).
5. Allen, J., Ferguson, G., Blaylock, N., Byron, D., Chambers, N., Dzikovska, M., Galescu, L., Swift, M.: Chester: Towards a personal medication advisor. J. Biomed. Inform. 39, 500–513 (2006).
6. Fadhil, A.: A Conversational Interface to Improve Medication Adherence: Towards AI Support in Patient's Treatment. (2018).
7. Zhang, Z., Takanobu, R., Zhu, Q., Huang, M., Zhu, X.: Recent advances and challenges in task-oriented dialog systems. Sci. China Technol. Sci. 1–17 (2020).
8. Li, X., Chen, Y.-N., Li, L., Gao, J., Celikyilmaz, A.: Investigation of language understanding impact for reinforcement learning based dialogue systems. arXiv Prepr. arXiv1703.07055. (2017).
9. Dusenberry, M.W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., Dai, A.M.: Analyzing the role of model uncertainty for electronic health records. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 204–213 (2020).
10. Louvan, S., Magnini, B.: Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey. arXiv Prepr. arXiv2011.00564. (2020).
11. Yao, K., Zweig, G., Hwang, M.-Y., Shi, Y., Yu, D.: Recurrent neural networks for language understanding. In: Interspeech. pp. 2524–2528 (2013).
12. Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G., Shi, Y.: Spoken language understanding using long short-term memory neural networks. In: 2014 IEEE Spoken Language Technology Workshop (SLT). pp. 189–194. IEEE (2014).
13. Gal, Y.: Uncertainty in deep learning. Univ. Cambridge. 1, 4 (2016).
14. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K.: Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. 28, 231–237 (2019).
15. Gauerhof, L., Munk, P., Burton, S.: Structuring validation targets of a machine learning function applied to automated driving. In: International Conference on Computer Safety, Reliability, and Security. pp. 45–58. Springer (2018).
16. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016).
17. Vasudevan, V.T., Sethy, A., Ghias, A.R.: Towards better confidence estimation for neural models. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7335–7339. IEEE (2019).
18. Bohus, D., Rudnicky, A.: Sorry and I Didn't Catch That!-An Investigation of Non-understanding Errors and Recovery Strategies. In: Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. pp. 128–143 (2005).

19. Aftab, H., Shah, S.H.H., Habli, I.: Classification of Failures in the Perception of Conversational Agents (CAs) and Their Implications on Patient Safety. Studies in health technology and informatics. 281, 659–663 (2021).
20. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. Adv. Neural Inf. Process. Syst. 29, 1019–1027 (2016).
21. Zhang, L., Zhang, L.: An Ensemble Deep Active Learning Method for Intent Classification. In: Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence. pp. 107–111 (2019).
22. van der Westhuizen, J., Lasenby, J.: Bayesian LSTMs in medicine. arXiv Prepr. arXiv1706.01242. (2017).
23. Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., van der Lugt, A., de Bruijne, M.: Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. pp. 32–41. Springer (2020).
24. Ghoshal, B., Tucker, A., Sanghera, B., Wong, W.L.: Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). pp. 318–324. IEEE (2019).
25. Gautam, V., Gheraibia, Y., Alexander, R., Hawkins, R.D.: Runtime Decision Making Under Uncertainty in Autonomous Vehicles. In: Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2021). CEUR Workshop Proceedings (2021).
26. Larson, S., Mahendran, A., Peper, J.J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J.K., Leach, K., Laurenzano, M.A., Tang, L.: An evaluation dataset for intent classification and out-of-scope prediction. arXiv Prepr. arXiv1909.02027. (2019).
27. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. arXiv Prepr. arXiv1802.10501. (2018).
28. Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., Willke, T.L.: Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 550–564 (2018).
29. Kim, J.-K., Kim, Y.-B.: Joint learning of domain classification and out-of-domain detection with dynamic class weighting for satisficing false acceptance rates. arXiv Prepr. arXiv1807.00072. (2018).
30. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. arXiv Prepr. arXiv1511.06349. (2015).
31. Zheng, Y., Chen, G., Huang, M.: Out-of-domain detection for natural language understanding in dialog systems. IEEE/ACM Trans. Audio, Speech, Lang. Process. 28, 1198–1209 (2020).
32. Munsch, N., Martin, A., Gruarin, S., Nateqi, J., Abdarahmane, I., Weingartner-Ortner, R., Knapp, B.: Diagnostic Accuracy of Web-Based COVID-19 Symptom Checkers: Comparison Study. J. Med. Internet Res. 22, e21299 (2020).
33. Zwaan, L., Hautz, W.E.: Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key, (2019).
34. Mooney, P.: Medical Speech, Transcription, and Intent, https://www.kaggle.com/paultimothymooney/medical-speech-transcription-and-intent, last accessed 2021/04/20.
35. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? arXiv Prepr. arXiv1703.04977. (2017).

36. Machin, M., Guiochet, J., Waeselynck, H., Blanquart, J.P., Roy, M., Masson, L.: SMOF: A Safety Monitoring Framework for Autonomous Systems. IEEE Trans. Syst. Man, Cybern. Syst. 48, 702–715 (2018).
37. Bickmore, T., Trinh, H., Asadi, R., Olafsson, S.: Safety first: conversational agents for health care. In: Studies in Conversational UX Design. pp. 33–57. Springer (2018).
38. Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., Reynolds, N.: Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Heal. care informatics. 26, (2019).