Delft University of Technology

NLICE: Synthetic Medical Record Generation for Effective Primary Healthcare Differential Diagnosis

Al-Ars, Zaid; Agba, Obinna ; Guo, Zhuoran ; Boerkamp, Christiaan; Jaber, Ziyaad ; Jaber, Tareq

Citation (APA)
Al-Ars, Z., Agba, O., Guo, Z., Boerkamp, C., Jaber, Z., & Jaber, T. (2023). NLICE: Synthetic Medical Record Generation for Effective Primary Healthcare Differential Diagnosis. In C. Ceballos (Ed.), *Proceedings of the 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 397-402). (Proceedings - 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering, BIBE 2023). IEEE. https://doi.org/10.1109/BIBE60311.2023.00071

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

NLICE: Synthetic Medical Record Generation for Effective Primary Healthcare Differential Diagnosis

Zaid Al-Ars[1]   Obinna Agba[1]   Zhuoran Guo[1]   Christiaan Boerkamp[1]   Ziyaad Jaber[2]   Tareq Jaber[2]

[1]Accelerated Big Data Systems
Delft University of Technology
Delft, The Netherlands

[2]Medvice Digital Health
Sint Janssingel 92, 5211DA
's-Hertogenbosch, The Netherlands

Abstract—This paper offers a systematic method for creating medical knowledge-grounded patient records for use in activities involving differential diagnosis. Additionally, an assessment of machine learning models that can differentiate between various conditions based on given symptoms is also provided. We use a public disease-symptom data source called SymCat in combination with Synthea to construct the patients records. In order to increase the expressive nature of the synthetic data, we use a medically-standardized symptom modeling method called NLICE to augment the synthetic data with additional contextual information for each condition. In addition, Naive Bayes and Random Forest models are evaluated and compared on the synthetic data. The paper shows how to successfully construct SymCat-based and NLICE-based datasets. We also show results for the effectiveness of using the datasets to train predictive disease models. The SymCat-based dataset is able to train a Naive Bayes and Random Forest model yielding a 58.8% and 57.1% Top-1 accuracy score, respectively. In contrast, the NLICE-based dataset improves the results, with a Top-1 accuracy of 82.0% and Top-5 accuracy values of more than 90% for both models. Our proposed data generation approach solves a major barrier to the application of artificial intelligence methods in the healthcare domain. Our novel NLICE symptom modeling approach addresses the incomplete and insufficient information problem in the current binary symptom representation approach.

Index Terms—medical records, synthetic data, differential diagnosis, machine learning

I. INTRODUCTION

The process of differential diagnosis is defined as: *The distinguishing of a disease or condition from others presenting with similar signs and symptoms*. In the context of primary healthcare, differential diagnosis describes the process by which a doctor (or other medical practitioners) deduces a possible set of conditions which might be responsible for the symptoms presented by a patient. As a result, the doctor determines the next course of action, such as drug prescription, further testing to rule out disease alternatives, etc. However, despite the ubiquity and importance of this process in the encounter between patients and doctors, it is by no means an easy task. A study [1] has shown that in the United States, 1 in 20 outpatient visits is misdiagnosed. Even in cases where a misdiagnosis is not harmful to the patient's health, extra time and financial resources are spent while arriving at the correct diagnosis. This places an extra burden on the patient, the doctor and the healthcare system as a whole. Because

of advances in machine learning (ML) techniques [2] and computational capabilities [3], it should be possible to train ML models capable of supporting the task of obtaining a differential diagnosis given a patient's symptoms. Such models can then be integrated in assistive tools which would, at the very least, confirm the doctor's initial differential diagnosis or suggest possible conditions which might have been overlooked [4].

However, a limitation in the field of differential diagnosis is the absence of sizable and medically accurate public datasets, due to the sensitive nature of medical information [5] and the potential risk of data leakage [6]. Recent research on automating diagnosis tasks has explored creating synthetic patient records or real-world datasets from free, online symptom checkers. Two real-world datasets based on patient self-reports and conversations in a Chinese healthcare online medical forum are Muzhi [7] and Dxy [8]. However, due to the limited number of reported symptoms and conditions, these real-world datasets cannot be widely utilized in ML models. One possible approach for expanding the scope of symptoms and conditions is to construct synthetic patient records based on the relationships between conditions and symptoms. Published work [9] built synthetic patient datasets by using the SymCat symptom-disease database. Although these real-world datasets and synthetic datasets have the possibility of evaluating the performance of ML models in the automatic diagnosis task, the medical correctness of those datasets cannot be guaranteed because probabilities and statistics used in symptom-condition databases lack professional and reliable medical knowledge. Moreover, the representation of symptoms in current medical records is trivial and incomplete, which results in simplistic symptom-condition assignments (e.g. binary or one-hot vector encoding). This illustrates the need for professional and medically correct datasets in the differential diagnosis research area.

This paper presents a systematic method to simulate medically-correct and highly-expressive patient records, which integrates a medically standardized symptom modeling approach called NLICE (pronounced as */en-lais/*). NLICE symptom modeling is a novel way to enhance the symptom representation, which is therefore useful for identifying diseases that have similar symptoms. Furthermore, in collaboration with medical experts, medically correct symptom-condition

statistics are provided to increase the accuracy of simulated patient records. We generate two types of datasets: one built with only SymCat symptom-condition database, and another with augmented NLICE information. The SymCat dataset includes a total of 5 million records, covering 801 distinct conditions each with 376 potential symptoms. The NLICE dataset uses statistics provided by medical experts, and includes 55 conditions each with 137 potential symptoms. In this way, we can bridge the gap between limited available patient records and data-driven healthcare methodologies.

II. METHODS

In this paper, we use SymCat [10], a public symptom-condition data source, in combination with Synthea to generate the data used subsequently for the analysis. The following subsections provide more details regarding the selected data source and generator.

A. Synthea

Synthea [11], a synthetic patient population simulator, allows for the generation of *realistic* patient medical records. To avoid privacy concerns, the generator was developed relying on publicly available medical information and health statistics. The project is also fully open-source with a permissive license which allows prospective users to modify the codebase to suit target applications. In its earliest version, Synthea modeled conditions ranked as the top 10 causes of visits to a primary healthcare provider and the top 10 conditions according to the "*years of life lost*" metric. Since this initial version, support has been added for more conditions with many of these contributed by its active community. Synthea's expressive generic module framework provides means to encode a disease-symptom probabilistic data source into Synthea compatible modules.

B. SymCat

SymCat is *a disease calculator that uses hundreds of thousands of patient records to estimate the probability of disease* [10]. It provides an interface where users can supply information about the symptoms being experienced and receive a differential diagnosis. In addition, SymCat provides a conditions and symptoms directory. This knowledge base which is publicly available on SymCat's website[1] provides probabilistic relationships between 474 symptoms and 801 conditions.

A more formal description of SymCat's data is given below:

- A list of conditions C is provided.
- For each condition C_i, the age based odds $Pr(c = C_i|a = A_j)$ for each age group A_j are provided.
- Also, for each condition, the gender based odds $Pr(c = C_i|g = G_k)$ are provided given that $G_k \in \{male, female\}$.
- Race based odds $Pr(c = C_i|r = R_l)$ for each race group R_l for each condition are also provided.

[1]A scrapped CSV version of this data was provided by Alexis Smirnov of https://www.dialogue.co/en and is publicly available at https://github.com/teliov/SymCat-to-synthea

- For each condition C_i, a set of symptoms S^i which might be presented for that condition along with the probability that the symptom is presented $Pr(S^i_m|c = C_i)$ where $S^i_m \in S^i$ is also provided.
- Additionally, for each symptom S_i, age based odds $Pr(s = S_i|a = A_j)$, race based odds $Pr(s = S_i|r = R_l)$ and gender based odds $Pr(s = S_i|g = G_k)$ are also provided.

Combining SymCat and Synthea: We developed a Python application[2] in collaboration with Arsène Fansi Tchango[3] of the MILA[4] research institute in Quebec. The application parses the CSV SymCat data and generates Synthea compatible modules. When generating the Synthea modules, the probability $Pr(c = C_i|a = A_j, g = G_k, r = R_l)$ is first determined using data provided in SymCat. In order to simplify the generation process, a conditional independence is assumed between the patient's age, gender and race given the patient's condition. Also assuming that all conditions have the same prior and with a repeated application of Bayes law we obtain:

$$Pr(c = C_i|a = A_j, g = G_k, r = R_l) = Pr(c = C_i|a = A_j)$$
$$\times Pr(c = C_i|g = G_k) \times Pr(c = C_i|r = R_l) \tag{1}$$

This gives the probability with which the Synthea generator will allow a patient with age A_j, gender G_k and race R_l to contract the disease or condition C_i. Once a patient record has been generated by Synthea, the presented symptoms are simply picked based on the probability $Pr(S^i_m|c = C_i)$ provided by the SymCat data source.

It is worth noting that the conditional independence assumption in data generation deviates from what is obtainable in practice. It is not unusual to have a more complex relationship between the patient demography and the contracted condition as well as the expressed symptoms [12]. Nonetheless, with the data available, this was a reasonable approximation to make.

C. NLICE database

The SymCat-based database mentioned above offers a machine-usable version of patient medical records: A binary value of 1 denotes the presence of a symptom, while a value of 0 denotes the absence of that symptom. However, we could provide more information about symptom characteristics without extra laboratory or diagnostic procedures. By doing so, we could better describe medical conditions and increase the effectiveness of differential diagnosis.

1) NLICE symptom modeling: Providing more characteristics in the symptom's expression can help make a better differential diagnosis in practice. These characteristics are summarized with the NLICE acronym, which is short for Nature, Location, Intensity, Chronology, and Excitation.

[2]https://github.com/teliov/SymCat-to-synthea
[3]https://github.com/afansi
[4]https://mila.quebec/en/mila/

We also use Synthea's expressive generic module framework to generate representative datasets that include this information as long as statistical relationships between these additional symptom characteristics and respective conditions are known.

Nature: Nature refers to the various ways a specific symptom can manifest itself [13]. For example, we use coughing as an illustration, which is a typical symptom connected to respiratory diseases. Modeling this symptom using a one-or-zero approach ignores the various ways coughing may occur. A dry cough could be experienced by some patients as opposed to others who may cough up mucous. These variations specify the type of cough and offer details that might make it simpler to differentiate between conditions that share similar symptoms.

Location: Location refers to the position on the body where a patient experiences a symptom. The location of a symptom can be a discriminating factor when making distinctions between conditions. For example, we consider a patient who reports experiencing abdominal pain. In terms of medical anatomy, the abdomen can be divided into upper and lower right quadrants as well as upper and lower left quadrants [14].

Intensity: The intensity of a symptom refers to the severity at which a symptom is experienced. The severity of the symptom presentation may clearly indicate the underlying illness. For instance, a common symptom of many ailments is pain. An appendicitis sufferer will commonly experience severe to moderate abdominal discomfort.

Chronology: Chronology encompasses three different concepts: 1. frequency, which refers to how frequently a symptom occurs, 2. duration, which refers to how long the symptom lasts, and 3. onset, which describes the time the patient first noticed the symptom. When determining a differential diagnosis, these characteristics may provide diagnostic value.

Excitation: Activities that patients engage in or situations patients are exposed to that activate or worsen the symptoms are referred to as excitation [15]. For instance, a patient may only experience heart pain while swimming. We could also distinguish conditions more precisely if we recorded excitation information.

2) NLICE data collection: An NLICE modeling strategy is not directly applicable for the data presented in SymCat. This is despite the fact that there are conditions that presented, for example, with *burning-abdominal* pain and others that presented with *sharp-abdominal* pain, thereby allowing SymCat to capture some elements of the NLICE technique (both these distinctions are made on the nature of the abdominal pain). However, not all symptoms that could support the NLICE strategy are covered by this approach. Therefore, we acquired data for a list of conditions from the medical literature as a proof of concept. This data was collected by our industry partners at Medvice, who are medical specialists. The conditions in our database were divided into ten groups for the purposes of data collection.

3) Combining NLICE and Synthea: We created Synthea-compatible modules and an adjusted Synthea generator[5] to generate the dataset, similar to the application we created for SymCat-based databases.

We assume conditional independence still holds in this formula, which yields:

$$Pr(c = C_i | a = A_j^i, g = G_k, r = R_l) = Pr(c = C_i | a = A_j^i)$$
$$\times Pr(c = C_i | g = G_k) \times Pr(c = C_i | r = R_l) \tag{2}$$

where A_j^i stands for the probability of age group j in condition i.

Subsequently, the Synthea generator picks the presented symptoms according to the probability $Pr(S_m^i | c = C_i)$ provided by the NLICE data source. At the same time, each presented symptom will be associated with eight NLICE features: $[symptom : nature : location : intensity : frequency : duration : onset : excitation]$. However, each condition does not need to contain all NLICE features. The adjusted Synthea generator displays only those conditions for which "NLICE" is known.

D. Selected ML models

For this problem, we select two popular ML models often employed in the medical domain: naive Bayes and random forest.

Naive Bayes: The Naive Bayes model is the model we use as a baseline, given the probabilistic problem formulation. The conditional independence assumption used in the data generation process is the same assumption made by the Naive Bayes algorithm. Despite this strong assumption, Naive Bayes still achieves reasonable accuracy [16, 17]. Also, this model allows for the most suitable probability distribution function to be used for each feature. This flexibility allows us to assume that the patient's age is distributed according to a Gaussian distribution, the patient's race assumes a categorical probability distribution and the gender along with all the symptoms (which take on values of 0 or 1) assumes a Bernoulli distribution.

Random Forest: The suitability of a Random Forest model is also evaluated. Random Forest is an ensemble method that uses bagging to combine predictions from a collection of decision trees each trained on a random subset of features [18]. It has been shown to be very robust to noise and also avoids the over-fitting commonly associated with single decision trees [19]. With very few parameters for optimization and the nature of our problem (decision-making), this model is also a good alternative.

III. RESULTS

A. Evaluation metric scores

Both the Naive Bayes and Random Forest models were trained and evaluated on the baseline data. Three metrics

---

[5]https://github.com/guozhuoran918/NLICE

were selected on which the models would be evaluated: Top-1 accuracy, precision and Top-5 accuracy. Tab. I shows the results of both models using the selected evaluation metrics in SymCat and NLICE datasets.

Tab. I shows that the Naive Bayes model slightly outperforms Random Forest in the SymCat-based dataset, while Random Forest performs better in the NLICE-based dataset. We also notice the relatively low SymCat Top-1 and Top-5 accuracy scores compared to NLICE which underscores the diagnostic value of augmenting the dataset with extra symptom characteristics. This increase in accuracy can be attributed to a number of factors. One is the similarity in symptom expression in SymCat for similar conditions, which makes an accurate diagnosis very difficult if not impossible for ML models. Another factor is the number of symptoms expressed per condition. As mentioned earlier, the baseline dataset is allowed to have patients with conditions presenting only one symptom. In such a case, it would be impossible to distinguish conditions based on only one such symptom.

B. Qualitative evaluation

For a qualitative evaluation, we observe the predictions made by the models. Due to the large number of conditions in the dataset being evaluated, we select Asthma condition as a case study. Two evaluations are carried out. Tab. II shows the 5 conditions which the models most often misclassify the selected condition as. This is a sort of qualitative confusion matrix. We see from the results that in all cases for Asthma predictions, the *confusion* in the SymCat-based dataset is mostly due to other respiratory conditions. This reinforces the reasoning that the models are unable to distinguish between similar conditions based on symptoms alone in SymCat. In contrast when using NLICE, Asthma is not misclassified as other similar respiratory conditions. This indicates that modeling symptoms with NLICE characteristics can be helpful to distinguish the conditions that contain similar symptoms.

C. Posterior estimate comparisons

Besides the qualitative comparisons made in the previous discussion, we also observe the posterior estimates for each of the models. As Equation 1 states, given an instance, trained ML models can estimate the probability for each class that the instance belongs to this class. These probabilities are called the posterior probabilities, then this instance is classified to the class with the highest posterior probability [20].

Fig. 1 shows the experimental results in the SymCat dataset. Naive Bayes has very high Top-1 probability estimates (median of 0.97), when compared to the more moderate Random Forest (median of 0.57). While the median confidence level of Naive Bayes is reduced by 30.52%, and a 25.28% reduction for the Random Forest in the Top-5 case. Fig. 2 shows that same findings hold for the NLICE-based dataset. These results show that Naive Bayes assigns a higher probability estimate to its predicted class compared to Random Forest, which is a well-known behavior [21]. For both models in the SymCat dataset, however, when their predictions are wrong, the probability

estimates are much lower as can be seen in Figures 1c and 1d. Similarly, the probability estimates for Random Forest in the NLICE dataset are much lower compared to Top-n accurate as shown in Figures 2c and 2d. This behavior is reasonable because it would be desirable that the models are not very confident when they wrongly predict conditions. However, Naive Bayes yields an overconfident result (either 1 or 0) for almost every sample in the NLICE dataset. One possible explanation is that Naive Bayes fails to learn the complicated relationships between NLICE attributes because it inherently assumes conditional independence between NLICE attributes [22], which is not the case in our context.
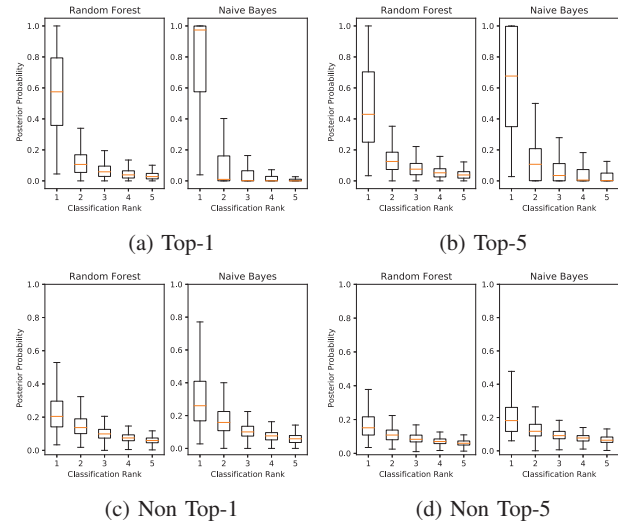

(a) Top-1 (b) Top-5

(c) Non Top-1 (d) Non Top-5

Fig. 1: Posterior probability estimate comparisons using Random Forest and Naive Bayes for the SymCat dataset

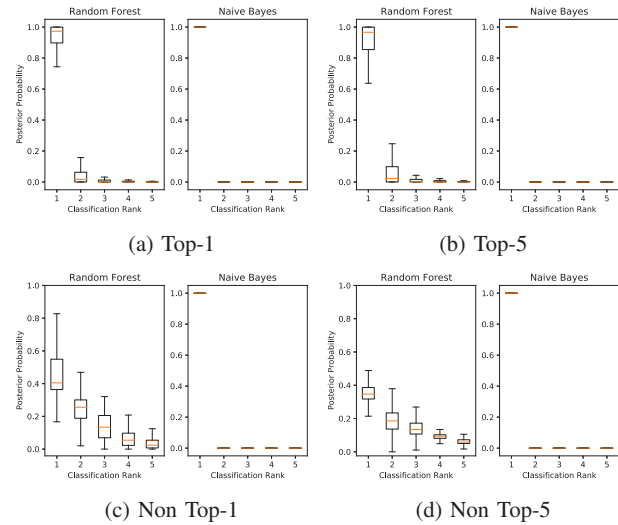(a) Top-1 (b) Top-5

(c) Non Top-1 (d) Non Top-5

Fig. 2: Posterior probability estimate comparisons using Random Forest and Naive Bayes for the NLICE dataset

TABLE I: Results of models on baseline data

Model	Top-1		Precision		Top-5	
	SymCat	NLICE	SymCat	NLICE	SymCat	NLICE
Naive Bayes	0.588	0.820	0.633	0.845	0.853	0.975
Random Forest	0.571	0.820	0.612	0.902	0.845	0.990

TABLE II: Top 5 commonly included predictions for Asthma

Condition	Naive Bayes		Random Forest	
	SymCat	NLICE	SymCat	NLICE
Asthma	Acute bronchospasm	Breast cancer	COPD	Breast cancer
	COPD	Otitis externa	Acute bronchospasm	Otitis externa
	ARDS	Lower urinary tract infection	ARDS	Lower urinary tract infection
	Croup	Otitis media acuta	Croup	Otitis media acuta
	Acute bronchiolitis	Atrophic gastritis	Acute bronchiolitis	Atrophic gastritis

D. Prediction confidence threshold

In a clinical setting, it is important to indicate when the predictions are not confident enough to make clinical decisions. As the results in the previous section indicate, models tend not to be confident when their predictions are wrong, which allows for the use of a confidence threshold in practice [23]. In such a setting, predicted conditions would be presented to the doctor only if the most accurate prediction exceeds the set confidence threshold. As an example, the confidence threshold is selected to be 25th, 50th and 75th percentile as well as the mean of the highest posterior estimate for the model's predictions.

Fig. 3 shows the application of these thresholds to the predictions of Random Forest for the SymCat dataset. We use the same cases mentioned in Section III-C: Top-1 accurate, Top-5 accurate, Non Top-1 accurate and Non Top-5 accurate. We calculate the 25th, 50th and 75th percentile as well as the mean of the Top-5 posterior estimate probabilities for each case as threshold values. Subsequently, those threshold values are used to re-evaluate the prediction results in the whole sample tests. In addition, the figure shows the percentage of conditions considered as confident depending on the threshold being used. It also shows what percentage of the confident predictions are accurate. Higher threshold values result in fewer *confident* predictions but also higher accuracy. As can be seen in Fig. 3a for Top-1 accurate predictions, using the 25th percentile as a threshold admits approximately half of the predictions as confident (approx. 50% of diagnosed conditions) but has a Top-1 accuracy of 85% and a Top-5 accuracy of 98%. Taking the 25th percentile of the Non Top-1 accurate case, the model obtains a 63% Top-1 accuracy and a 89% Top-5 accuracy and it admits 88% of the predictions. In a deployed environment, such a threshold value might be considered reasonable. Predictions below the threshold would not typically be presented to the doctor.

Fig. 4 represents the Random Forest results for the NLICE dataset. The figure shows a significant difference compared to the SymCat dataset: NLICE Top-1 threshold results in Fig. 4a and Top-5 threshold results in Fig. 4b show that increasing the threshold values reduces the number of confident predictions. Non Top-n threshold results in Fig. 4c and Fig. 4d slightly reduce Top-1 accuracy as the threshold confidence increases.

However, most of the predicted conditions by both models can achieve around 100% accuracy. And the diagnosed percentages in the NLICE dataset are higher than the SymCat dataset, indicating that NLICE models are more confident about their predictions than SymCat models. Applying Non Top-n threshold values in NLICE models admits around 90% of predictions in the 75th percentile, which means that most Top-1 conditions are assigned a 75th percentile probability in all conditions.
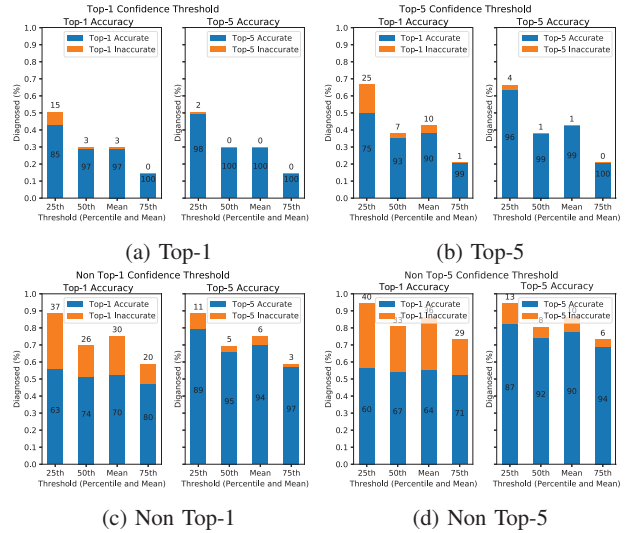

(a) Top-1   (b) Top-5

(c) Non Top-1   (d) Non Top-5

Fig. 3: Prediction confidence threshold using Random Forest for the SymCat dataset

IV. CONCLUSION

In this paper, we proposed a systematic approach to constructing synthetic patient records. As has been stated earlier, symptom information is not always enough to make a proper diagnosis, hence additional information regarding the nature, location, chronology, etc, of the symptom, can be gathered from medical literature to increase the available information to the models. We collected additional symptom information for selected conditions. For standardization of the symptom expression, we proposed a novel symptom modeling approach called NLICE and integrated this symptom

(a) Top-1   (b) Top-5
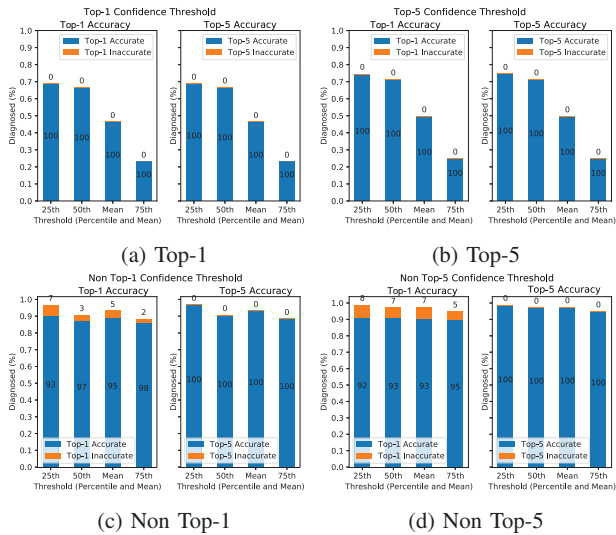
(c) Non Top-1   (d) Non Top-5

Fig. 4: Prediction confidence threshold using Random Forest for the NLICE dataset

modeling approach with the Synthea simulator. The analysis demonstrated the suitability of these datasets for using ML models (e.g., Naive Bayes and Random Forest) to the task of estimating a differential diagnosis given a patient's symptoms and demography. We trained Naive Bayes and Random Forest models in both datasets and demonstrated their suitability. The feasibility of using a confidence threshold to filter out the most likely incorrect predictions in a deployed setting was also demonstrated. Qualitatively, models trained in the SymCat-based dataset struggle to distinguish conditions that share similar symptoms. In contrast, this problem did not occur in the models trained in the NLICE-based dataset, which also highlights the importance of introducing additional features and information on symptoms.

We are working on a couple of future research directions. For example, given the complexity of symptom-disease relationships, we need to explore more powerful and robust ML modeling techniques like autoencoders [24]. Another potential direction for research would be using more sophisticated metrics since very dissimilar conditions can both present similar symptoms and hence the similarity to the actual condition becomes inadequate.

Acknowledgment

References

[1] H. Singh, A. N. Meyer, and E. J. Thomas, "The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations," *BMJ Qual Saf*, vol. 23, no. 9, pp. 727–731, 2014.

[2] B. Zhu, Z. Al-Ars, and H. P. Hofstee, "Nasb: Neural architecture search for binary convolutional neural networks," in *2020 International joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[3] J. Hoozemans, J. Peltenburg, F. Nonnemacher, A. Hadnagy, Z. Al-Ars, and H. P. Hofstee, "Fpga acceleration for big data analytics: Challenges and opportunities," *IEEE Circuits and Systems Magazine*, vol. 21, no. 2, pp. 30–47, 2021.

[4] O. Kostopoulou, A. Rosen, T. Round, E. Wright, A. Douiri, and B. Delaney, "Early diagnostic suggestions improve accuracy of gps: A randomised controlled trial using computer-simulated patients," *British Journal of General Practice*, vol. 65, 12 2014.

[5] S. Zwaard, H.-J. Boele, H. Alers, C. Strydis, C. Lew-Williams, and Z. Al-Ars, "Privacy-preserving object detection & localization using distributed machine learning: A case study of infant eyeblink conditioning," *arXiv:2010.07259*, 2020.

[6] D. Enthoven and Z. Al-Ars, "An overview of federated deep learning privacy attacks and defensive strategies," *Federated Learning Systems*, pp. 173–196, 2021.

[7] Z. Wei, Q. Liu, B. Peng, H. Tou, T. Chen, X. Huang, K.-f. Wong, and X. Dai, "Task-oriented dialogue system for automatic diagnosis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 201–207. [Online]. Available: https://aclanthology.org/P18-2033

[8] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, and L. Lin, "End-to-end knowledge-routed relational dialogue system for automatic diagnosis," 2019. [Online]. Available: https://arxiv.org/abs/1901.10623

[9] K. Liao, Q. Liu, Z. Wei, B. Peng, Q. Chen, W. Sun, and X. Huang, "Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning," *arXiv preprint arXiv:2004.14254*, 2020.

[10] A. R. Inc., "Symcat: Symptom-based, computer assisted triage," http://www.symcat.com/, 2020.

[11] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 230–238, 2018.

[12] Y. Cheng, A. Kanaya, M. Araneta, S. Saydah, H. Kahn, E. Gregg, W. Fujimoto, and G. Imperatore, "Prevalence of diabetes by race and ethnicity in the united states, 2011-2016," *JAMA*, vol. 322, no. 24, pp. 2389–2398, 2019.

[13] L. S. Franco, D. F. Shanahan, and R. A. Fuller, "A review of the benefits of nature experiences: More than meets the eye," *International journal of environmental research and public health*, vol. 14, no. 8, p. 864, 2017.

[14] M. J. Hepburn, D. P. Dooley, S. L. Fraser, B. K. Purcell, T. M. Ferguson, and L. L. Horvath, "An examination of the transmissibility and clinical utility of auscultation of bowel sounds in all four abdominal quadrants," *Journal of clinical gastroenterology*, vol. 38, no. 3, pp. 298–299, 2004.

[15] J. Lisman, "Excitation, inhibition, local oscillations, or large-scale loops: what causes the symptoms of schizophrenia?" *Current opinion in neurobiology*, vol. 22, no. 3, pp. 537–544, 2012.

[16] M. P. Pedro Domingos, "Beyond independence:conditions for the optimality of the simple bayesian classifier," *Machine Learning*, vol. 29, pp. 103–130, 1997.

[17] H. Zhang, "The optimality of naive bayes," in *FLAIRS Conference*, 2004.

[18] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE TRANSACTIONS ON GEOSCIENCE ELECTRONICS*, vol. GE-15, no. 3, pp. 142–148, Jul. 1997.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] V. Pronk, S. V. R. Gutta, and W. F. J. Verhaegh, "Incorporating confidence in a naive bayesian classifier," in *User Modeling 2005*, L. Ardissono, P. Brna, and A. Mitrovic, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 317–326.

[21] P. N. Bennet, "Assessing the calibration of naive bayes' posterior estimates," Carnegie Mellon University, Tech. Rep., Sep. 2000.

[22] D. A. Moore and P. J. Healy, "The trouble with overconfidence." *Psychological review*, vol. 115, no. 2, p. 502, 2008.

[23] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] R. Miotto, L. Li, and J. T. Dudley, "Deep learning to predict patient future diseases from the electronic health records," in *Advances in Information Retrieval*. Cham: Springer International Publishing, 2016, pp. 768–774.