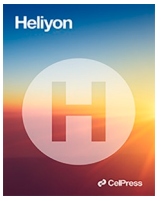
Heliyon 10 (2024) e26770

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Robust diagnosis recommendation system for Primary Care Telemedicine using long short-term memory multi-class sequence classification

Patrick Essay *, Ajaykumar Rajasekharan

Teladoc Health, Inc, 1875 Lawrence St, Denver, CO, 80202, USA

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Recurrent neural networks
Recommender systems
Clinical decision support
Electronic health records

ABSTRACT

Background: Telemedicine offers opportunity for robust diagnoses recommendations to support healthcare providers intra-consultation in a way that does not limit providers ability to explore diagnostic codes and make the most appropriate selection for each consultation.
Objective: The objective of this work was to develop a recommendation system for ICD-10 coding using multiclass sequence classification and deep learning. The recommendations are intended to support telemedicine clinicians in making timely and appropriate diagnosis selections. The recommendations allow clinicians to find and select the best diagnosis code much quicker and without leaving the telemedicine platform to search codes and code descriptions.
Methods: We developed an LSTM model for multi-class text sequence classification to make diagnosis recommendations. The LSTM recommender used text-based *symptoms, complaints,* and *consultation request reasons* as model inputs. Data were extracted from a live telemedicine platform which spans general medicine, dermatology, and mental health clinical specialties. A popularity-based model was used for baseline comparison.
Results: Using over 2.8 MM telemedicine consultations during 2021 and 2022, our LSTM recommender average accuracy was 31.7%. LSTM recommender average coverage in the top 20 recommended diagnoses was 85.8% with an average personalization score of 0.87.
Conclusions: LSTM multi-class sequence classification recommends diagnoses specific to individual consultations, is retrainable on regular intervals, and could improve diagnoses recommendations such that providers require less time and resources searching for diagnosis codes. In addition, the LSTM recommender is robust enough to make recommendations across clinical specialties such as *general medicine, dermatology,* and *mental health.*

1. Introduction

The International Classification of Diseases (ICD) is globally recognized as a standardized diagnostic tool maintained by the World Health Organization. In the United States, the 10th revision (ICD-10) is used as a diagnosis coding system for medical record-keeping in conjunction with Current Procedural Terminology (CPT) codes for inpatient billing [1]. Telemedicine, unlike a traditional inpatient setting, may require clinicians to be responsible for ICD-10 selection rather than administrative or medical coding staff. Due to potential knowledge gaps with coding procedures, errors could be more frequent [2–4]. Additionally, this presents a shift from

* Corresponding author.
E-mail address: patrick.essay@teladochealth.com (P. Essay).

https://doi.org/10.1016/j.heliyon.2024.e26770
Received 11 January 2024; Received in revised form 12 February 2024; Accepted 20 February 2024
Available online 29 February 2024

administrative billing-focused coding in an inpatient setting to coding for telemedicine record-keeping, irrespective of financial reimbursement and insurance requirements.

Medical recommendation systems have been developed in various settings for different objectives [5]. Traditional, collaborative- and content-based filtering methods, machine learning, and deep learning models have served as clinical decision support tools by recommending lab orders, medications, and diagnosis codes for medical billing with variable success [6–9]. Recommender systems have also been leveraged for inpatient hospitalization outcomes prediction and treatment path decision-making [10–12]. In addition to minimization of errors, recommendation systems can improve coding efficiency, decrease search time, and have the potential to allow providers to utilize a single search platform rather than leveraging multiple tools for ICD search and coding.

Recommendation methodologies for ICD coding have also varied greatly [13]. Neural networks and attention-based models have been used for report classifications based on ICD codes. Natural language processing has also been used in combination with convolutional neural networks for diagnosis prediction based on open-text clinical notes [14]. To our knowledge, a diagnosis recommender system has not been developed for telemedicine consultations, and previous approaches may be insufficient to address the breadth of illnesses across general telemedicine or other clinical specialties.

The objective of this study was to develop a robust ICD-10 recommender system for a telemedicine platform performing general medicine, dermatology, and mental health consultations. Telemedicine platforms are uniquely positioned to employ sophisticated decision support systems. Our goal was to support clinicians making ICD-10 code selections by actively recommending diagnoses for individual telemedicine consultations. Our approach supports clinicians in such a way that does not limit clinician's ability to view and select the most appropriate diagnoses and likely would not promote over-reliance on the recommender system for diagnosis code selection [15].

2. Material and methods

2.1. Data source and preprocessing

Deidentified data were extracted from a national telehealth network in the United States from September of 2021 through September of 2022. All available adult (≥18 years old) network members across clinical specialties (i.e., general medicine, mental health, and dermatology) with at least one consultation were included in the analyses. Both initial and follow-up consultations for a single illness or multiple consultations for differing illnesses were included.

Telemedicine consultation data included unique consultations across general medicine, dermatology, and mental health specialties, each of which may include clinical providers with multiple subspecialties (Appendix Table 1). Data also included patient demographics (age and gender), symptoms, complaints, consultation request descriptions, and primary ICD-10 diagnosis codes for each consultation. Symptoms and complaints consisted of structured dropdown menu selections while consultation request descriptions were unstructured, open text data fields.

Data preprocessing and analyses were performed using Python Language Reference Version 3.10.6 (Python Software foundation, Wilmington, DE), the Pandas (v.0.23.4) [16], Seaborn (v.0.9.0) [17], Sci-kit Learn package (v.0.19) [18], and TensorFlow [19] libraries. Methods were in alignment with the TRIPOD statement for predictive modeling (Appendix Table 2) and recommended reporting guidelines for machine learning algorithms and medical artificial intelligence [20,21].

2.2. Long short-term memory multi-class sequence classification

We developed a deep learning long short-term memory (LSTM) model for multiclass sequence classification for primary diagnosis code recommendations. LSTM was selected primarily for the ability to retain meaningful information and remove irrelevant information through a forget gate. The model is also able to remember longer sequences of information and extract meaningful insight. Lastly, it is less prone to vanishing gradient problems [22].

All available features directly related to individual consultations were considered. Coarse grained features, such as patient characteristics, provider characteristics, or provider-member-ICD interactions were excluded. Our model input data, including *consultation request reason*, *clinical complaints*, and *clinical symptoms,* are directly related to each consultation individually and are recorded via the clinician platform prior to diagnosis selection.

Table 1
Patient characteristics and input features stratified by service specialty.

	General Medicine	Dermatology	Mental Health	Total
Consultations, n	2,671,345	45,478	108,749	2,825,572
Providers, n	3400	88	2973	6461
Patients, n	1,921,667	40,711	68,235	2,001,799
Age, median (IQR)	36.3 (20.5)	31.9 (16.3)	32.6 (14.7)	36.1 (20.2)
Gender, %female	63.6	61.2	67.1	63.7
Unique diagnoses, n	7035	705	381	7131
Consultation request reasons, n	418,130	10,784	12,144	439,655
Unique symptoms, n	715	167	393	729
Unique complaints, n	279	73	118	296
Mean symptoms per consult, n	2.05	1.10	1.82	2.02

Table 2
Model performance metrics.

	Popularity Model	LSTM Sequence Classification
Accuracy, %	3.05	31.7
Coverage (top 20), %	43.9	85.8
Personalization	0	0.87

Consultation request reason is written by the patient during the consultation scheduling process. It is open-ended text format and can be as simple as "earache" or consist of longer descriptions such as "cold stuffy runny nose cough sneezing". Clinical complaints and clinical symptoms are selected by the medical provider during the consultation prior to diagnosis selection. Complaints and symptoms are standardized, dropdown menu selections and include items such as depression, checkup, nasal congestion, abdominal pain, etc.

Both structured (dropdown menu) and unstructured (open text) data were tokenized and then padded to a maximum sequence length of 20 words (Fig. 1). Primary diagnosis codes were factorized for each consultation. Consultations may contain multiple diagnoses, but only primary diagnoses were considered.

The LSTM model included an embedding layer, LSTM layer with dropout of 0.2, and a dense connected layer to SoftMax output layer (Fig. 2). Tokenized feature sequences were input to the embedding layer with embedding vector length of 100. The number of classes in our multiclass classification approach was determined by the number of diagnoses in the training and testing sets. Thus, the activation function in the last LSTM dense layer outputs a probability for each factorized diagnosis code. For diagnosis codes to be included in the training set, each must have been used a minimum of two times during the training date range.

The LSTM model outputs a probability for each possible diagnosis code based on a given input sequence. Every consultation returned a list of probabilities whose length equaled the number of unique diagnosis codes. These probabilities were then used to rank the diagnosis codes for each consultation input sequence from most likely to least likely diagnosis (Appendix Fig. 1).

2.3. Training and testing

The LSTM model was trained and tested using a rolling 14 days of consultation data. A train-test split size of 0.30 was used. Due to the low number of feature sequences from the standardized dropdown menu selection options relative to open text inputs, model training epochs were limited to 10 to avoid overfitting. Results from each two-week period were then compiled and averaged across one full year of data. The Results section includes both compiled performance over one year and monthly performance, i.e., roughly two 14-day cycles each month.

On average, each train-test cycle contained 2500+ diagnoses and approximately 50,000 unique input sequences for 140,000+ consultations. The model was trained and tested for every two-week period during the full year of extracted data. This was done for two reasons: 1) to limit computational burden and 2) to minimize impact of seasonality. If implemented to a live telemedicine platform, recommendations would then be based on the previous 14 days of consultations rather than a full year. For example, this approach may help avoid over-recommendation of seasonal influenza diagnoses during summer months when heat related illnesses may be more prevalent and vice versa.

Both models were tested across all three clinical specialties combined and individually to evaluate the need for recommendations at a more granular level. For instance, if a specialty area consistently requires a small number of diagnosis codes, then the need for a recommender system is diminished. But if individual specialties still treat a broad range of diagnoses, the impact of the recommender system being robust is nontrivial.

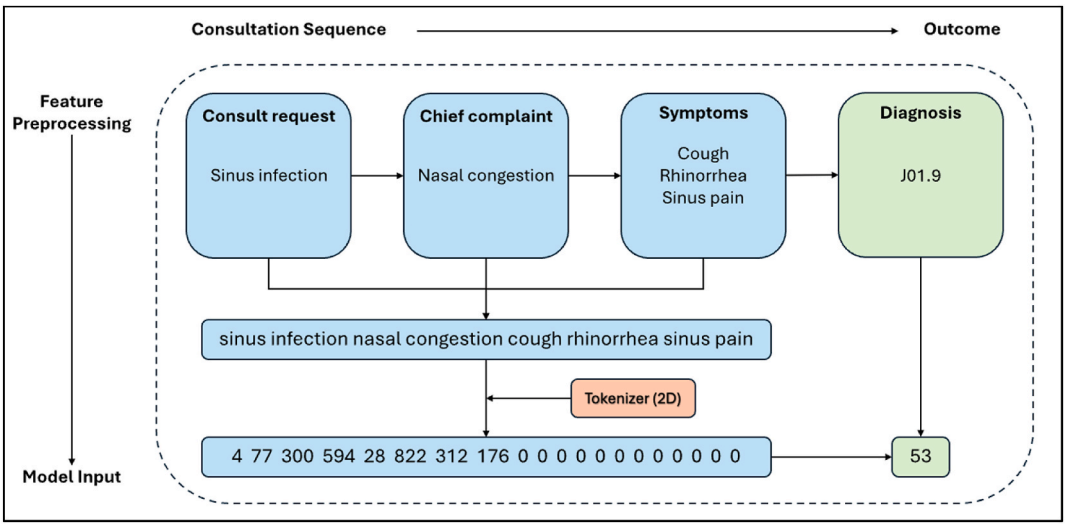


Fig. 1. Tokenization sequence of input features.

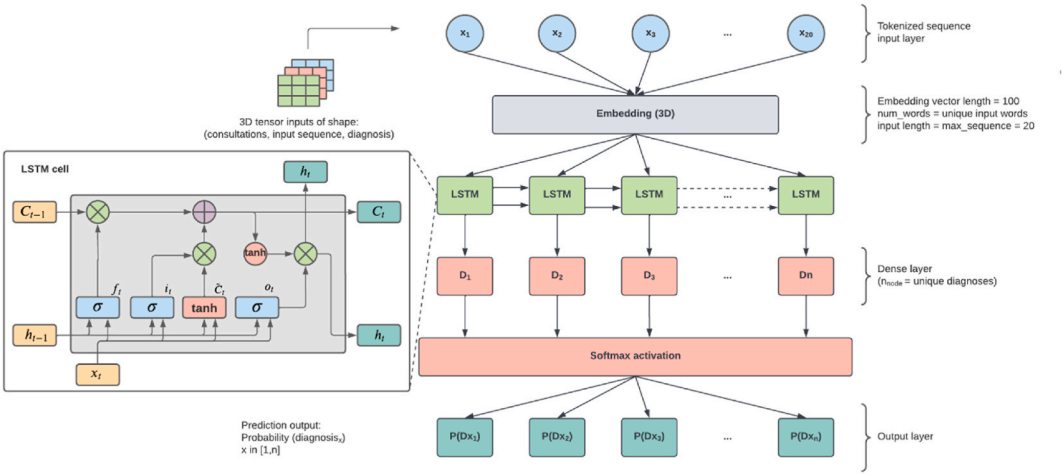


Fig. 2. LSTM model structure and input sequence.

2.4. Evaluation and baseline comparison

For baseline comparison, we created a popularity-based recommender model. The popularity model leveraged the frequency of primary diagnosis codes from all consultations across clinical specialties, combined and individually. Model training required that each ICD-10 code be selected in more than one consultation. Training data also required that providers had a minimum of two consultations which resulted in a primary diagnosis selection. Providers with only one consultation were excluded. Training and testing of the popularity model were performed across the same timeframe as the LSTM recommender system for results comparison.

Accuracy, coverage, and a personalization metric were calculated for evaluation and comparison of the baseline popularity model to the LSTM sequence classification model. Accuracy in this application refers to the percentage with which the correct diagnosis for a given consultation in the testing data was the first recommendation in the ranked list of model diagnoses output.

Coverage and personalization both used the top 20 recommended diagnoses for each consultation in test data. Coverage was calculated as the percentage of consultations in the testing data in which the correct diagnosis was present in the top 20 recommended diagnoses.

Personalization was calculated by finding the average cosine similarity measure (Eqn. (1)) for the top 20 diagnosis recommendations of each consultation where $k = \textit{cosine similarity}$ and x and y are row vectors.

$$k(x, y) = \frac{xy^{\mathrm{T}}}{\|x\|\|y\|} \qquad \text{Equation 1}$$

The personalization metric was then calculated as $1 - \textit{average cosine similarity}$ for the top 20 recommended diagnoses for each input sequence. Values closer to 1 suggest higher personalization across consultations and values closer to 0 (lower personalization) suggest that the same recommendations are being made for each consultation regardless of input sequence heterogeneity.

3. Results

The study dataset consisted of more than 2 million patients and 2.8 million consultations across three clinical specialties over one year (Table 1). The median age of members was 36 years. The data included 7131 unique ICD-10 diagnoses. Most diagnosis codes were from general medicine consultations with an average of two clinical symptoms per consultation.

A total of 264,561 (9.26%) consultations did not have a *consultation request reason.* All other input features for all consultations had no missing values (Appendix Table 3). Consultations with missing *consultation request reason* were not excluded. Our approach combines inputs into a single tokenized sequence string of values. Missing input data results in a potentially shorter input sequence but still allows for diagnosis recommendations.

3.1. Model performance comparison

Popularity model average accuracy (correct diagnosis was the first recommended diagnosis) was 3.02%, and LSTM model accuracy was 31.6% on average (Table 2) across all specialties combined. Model coverage where the correct diagnosis was present in the top 20 recommended diagnoses was 44.9% and 84.6% for the popularity model and LSTM model, respectively. The personalization score was 0.86 for the LSTM model meaning the difference in the top 20 recommended diagnoses for each consultation was very high. The popularity model personalization score was 0 because the same list of diagnosis codes was recommended for every test consultation, whereas the LSTM model was highly personalized based on the input sequences.

Model performance did not vary throughout the year (Fig. 3) suggesting seasonality has a minimal effect on the total patient population. Average performance metrics and figures were calculated using randomly selected consultations from the full dataset.

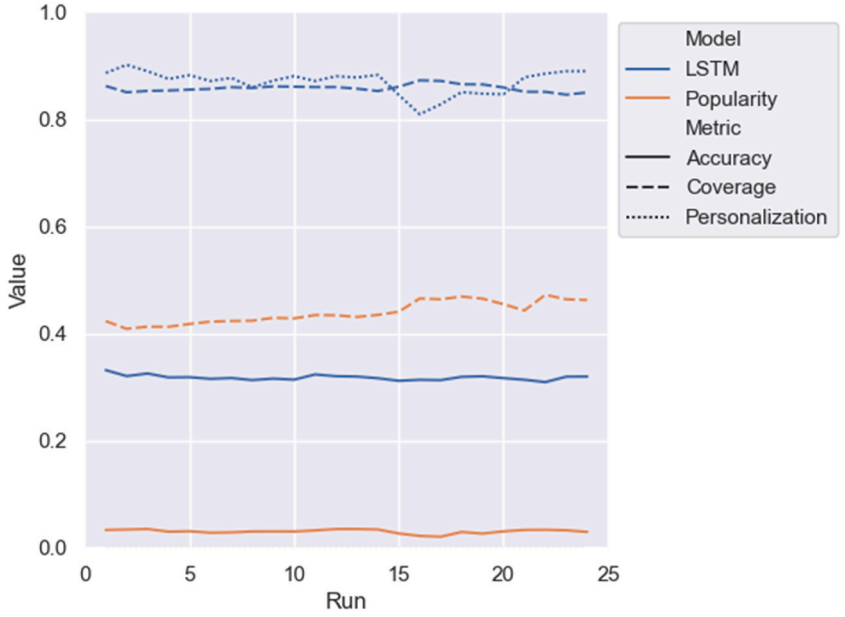


Fig. 3. LSTM and popularity model accuracy, coverage, and personalization for models trained and tested every two weeks through the full 1-year dataset.

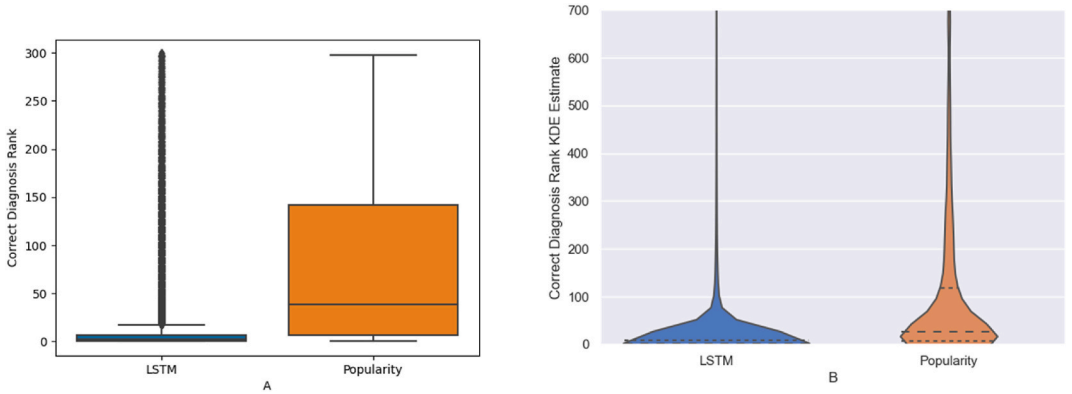


Fig. 4. A) Boxplot of the ranking of each correctly recommended diagnosis in testing consultations for both LSTM sequence classification model and popularity model, and B) violin plots illustrating kernel density estimates of the ranking of each correctly recommended diagnosis in testing consultations for both models.

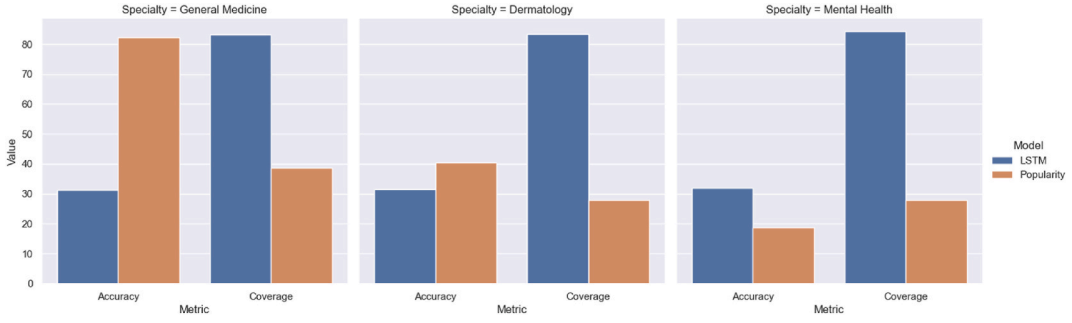


Fig. 5. Accuracy and coverage performance metrics for clinical specialties General Medicine, Mental Health, and Dermatology.

Similar to the coverage evaluation metric, the ranking of the correct diagnosis code for each model was higher for the LSTM model than the popularity model (Fig. 4) where ranking 0 was the top ranked diagnosis. Both the box plot (left) and violin plot (right) illustrate the correct diagnosis was more often and more likely to be ranked higher by the LSTM model. Qualitatively, the same was

true of other relevant diagnosis codes that were not selected. The popularity model was unable to provide different rankings for each consultation resulting in a ranking that only represents the most common consultations.

Model performance remained high across clinical specialties individually for the LSTM model as well and performance metrics were mixed for the popularity model (Fig. 5). Performance metrics are also shown in Appendix Table 4.

The performance of the LSTM model for individual specialties mirrored the performance when tested across all specialties combined. For the popularity model, however, there was an increase in performance across all three specialties, particularly general medicine. This is because the number of diagnoses treated within an individual specialty is smaller than the number of diagnoses treated across all specialties combined.

It is likely that the accuracy of the popularity model in general medicine of 82.3% is due to a very high number of cold and flu patients which is the most common general medicine illness (Appendix Fig. 2). Conversely, the popularity model for general medicine only has a coverage of 38.6% meaning the correct diagnosis is in the top 20 recommended diagnoses only 38.6% of the time. So, when patients schedule a general medicine visit for illnesses unrelated to cold and flu the popularity model performs worse while coverage of the LSTM model across all individual specialties and when specialties are combined remains above 80%.

4. Discussion

4.1. Performance and clinical relevance

Our approach using LSTM deep learning for multiclass text sequence classification to recommend clinical diagnoses performed well and far exceeded our baseline popularity modeling approach. Currently, an alpha-numeric ranking is used where three or more search characters are used to populate matching diagnosis codes and descriptions. Alpha-numeric diagnoses search functionality may be cumbersome and unspecific to individual consultations. In addition, the expansive number of possible ICD-10 codes makes searching for the correct diagnosis codes difficult. The popularity modeling approach is certainly an improvement over character search by recommending the most popular diagnoses over a specific timeframe. Yet, the LSTM model outperformed popularity modeling across all performance metrics.

Perhaps most importantly, the LSTM model makes recommendations specific to each consultation (Table 2). Telemedicine platforms allow for patient input *consultations request reasons* and clinician input *chief complaints* and *symptoms* prior to making a primary diagnosis selection. These text data are highly correlated and relevant to the consultation without expanding model inputs to include other patient-related factors such as age, weight, blood pressure, etc. which may or may not be present for each patient.

With the high personalization of multiclass classification, the model is robust enough to effectively make recommendations for multiple clinical specialties, subspecialties, and illness types (mental health, dermatology, general medicine, etc.). It is accurate enough to rank correct diagnoses high in listed recommendations (Figs. 4 and 5). And, the LSTM model can be trained on historical consultation data and generate recommendations in real time.

The LSTM model has several advantages for this application over other recommender systems such as collaborative and content-based filtering [23]. Traditional filtering methods could relate characteristics between clinicians (specialties), patients (physiology), diagnosis codes (illness types), and interactions between the three to make recommendations. However, in our case characteristic similarities and interactions between patients-providers-diagnoses are not strongly related to the correct diagnosis for a consultation. For example, clinicians do not necessarily select diagnoses based on preference. Alternatively, patient characteristics (such as age, body mass index, basic physiology) may not be related to their reason for telemedicine consultation.

4.2. Limitations and future work

While popularity modeling may be an improvement over alpha-numeric search, it has several drawbacks. The diagnosis codes that are highly recommended may become more popular over time as a result of being recommended [24–26]. This requires larger shifts in clinical practice or drastic changes in frequent diagnoses across entire public health domains to change the recommended diagnoses in a popularity-based model. The popularity model also does not account for new ICD diagnosis codes or changes to existing code definitions and descriptions that might affect the popularity of those codes.

The LSTM recommendation system outperformed the popularity model across all metrics. It is, however, limited in that it requires real-time input during a consultation to make diagnosis recommendations. It also has yet to be tested prospectively to evaluate the computational load and speed with which recommendations are returned on a live telemedicine platform. In addition, LSTM models minimize vanishing gradient issues but do not eliminate vanishing gradient entirely [22]. They require high memory bandwidth for training due to the relative complexity within each layer which might inhibit implementation of the model in production environments. Lastly, they are prone to overfitting. The standardized lists which were used for clinical complaints and clinical symptoms may exacerbate potential overfitting of the model if not trained and tested across heterogeneous groups.

Future work will include prospective validation and testing. The LSTM recommendation system should allow providers to search more effectively without leaving the platform to identify ICD-10 codes of interest. It should shorten the amount of time spent searching for codes by actively populating recommendations as providers are typing and should not limit accessibility to any ICD-10 codes in the search results.

5. Conclusion

Telemedicine platforms can leverage sequential clinical data intra-consultation to make accurate diagnosis recommendations using LSTM multiclass sequence classification. Recommending diagnoses dynamically may allow for robust and actionable clinical decision

support in real time. Diagnosis recommender systems can improve consultation efficiency and potentially minimize coding errors.

Summary table

What was already known on the topic.

- Recommender systems have been used in clinical applications for specific clinical objectives to varying degrees of success.
- Recommender systems can serve to minimize medical errors, improve clinical process efficiency, and for system-level, provider-level, and patient-level outcomes prediction.
- Recommender systems traditionally leverage interactions between entities (provider-patient) and/or metadata between study subjects (similar patient-patient or provider-provider) to make recommendations.

What this study added to our knowledge:

- We developed a model that makes diagnosis code recommendations for telemedicine consultations across broad, general telemedicine and specific clinical specialties.
- Our model leverages patient input data (reason for visit), symptoms and complaints as open text to make recommendations.
- We illustrated that our model makes diagnosis recommendations with a high enough degree of accuracy to assist clinicians in timely diagnosis selection during a remote, telemedicine consultation without assistance from other support staff.
- Multi-class sequence classification of free text using an LSTM model sufficiently captured information related to final patient diagnosis.

Data availability statement

Due to privacy and HIPAA data used in this study are unable to be made available.

CRediT authorship contribution statement

Patrick Essay: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Ajaykumar Rajasekharan:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Patrick Essay reports financial support was provided by Teladoc Health Inc. Patrick Essay reports a relationship with Teladoc Health Inc that includes: employment. Ajaykumar Rajasekharan reports financial support was provided by Teladoc Health Inc. Ajaykumar Rajasekharan reports a relationship with Teladoc Health Inc that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Stefanie Painter, DHEd contributed to editing the manuscript.

Appendix

Table 1
Total number (%) of provider roles in extracted data as percentage of consultations.

Provider roles (consultations)	Providers, n	Consultations, n
DOCTOR	3354	2,660,010
DERMATOLOGIST	88	45,478
COUNSELOR	1246	37,638
SOCIALWORKER	990	27,525
PSYCHIATRIST	279	31,026
THERAPIST	278	8364
PSYCHOLOGIST	179	4140
NURSEPRACTITIONER	49	11,097
COUNSELORADDICT	3	56
PHYSICIANASSISTANT	2	238

Table 2
TRIPOD Checklist.

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, , predictors, outcome, statistical analysis, results, and conclusions.	3
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	4
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
	5b	Describe eligibility criteria for participants.	5
	5c	Give details of treatments received, if relevant.	5
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	5
	6b	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	6-8
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	6-8
Sample size	8	Explain how the study size was arrived at.	5
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	NA
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	6-8
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6-8
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	6-8
Risk groups	11	Provide details on how risk groups were created, if done.	5
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8,18
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8,18
Model development	14a	Specify the number of participants and outcome events in each analysis.	8, 18
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	9, 19
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	8,9
	15b	Explain how to the use the prediction model.	9, 10
Model performance	16	Report performance measures (with CIs) for the prediction model.	19-21
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	12
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	10-13
Implications	20	Discuss the potential clinical use of the model and implications for future research.	10-13
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	included
Funding	22	Give the source of funding and the role of the funders for the present study.	Included

Consultation	Diagnosis 1	Diagnosis 2	Diagnosis 3	...	Diagnosis n
0	5.95E-05	1.76E-05	0.00012267	...	0.00058292
1	3.33E-05	3.96E-05	9.67E-05	...	1.17E-05
2	2.26E-05	1.00E-05	3.84E-05	...	2.96E-05
...	9.31E-05	5.69E-05	5.23E-05	...	3.46E-06
n	9.31E-05	5.69E-05	5.23E-05	...	3.46E-06

↓

Consultation	Ranking 1	Ranking 2	Ranking 3	...	n
0	Dx189	Dx180	Dx201	...	Dx182
1	Dx 369	Dx475	Dx374	...	Dx476
2	Dx 189	Dx180	Dx187	...	Dx201
...	Dx 447	Dx320	Dx309	...	Dx30
n	Dx 447	Dx320	Dx309	...	Dx30

Fig. 1. Example illustration of LSTM model output where each possible diagnosis is assigned a probability for each consultation in testing data. Diagnosis codes are then ranked based on probability from most to least likely diagnosis for a given consultation.

Table 3
Total percentage missingness of input features and primary diagnosis codes.

Input Feature	Missing, %
Clinical symptoms	0
Clinical complaints	0
Consultation request reasons for visit	9.26

Table 4
Model performance metrics stratified by individual clinical specialties.

	Popularity			LSTM		
	Gen. Med.	Mental Health	Dermatology	Gen. Med.	Mental Health	Dermatology
Accuracy, %	82.3	18.7	40.4	31.3	31.9	31.5
Coverage, %	38.6	68.7	27.8	83.1	84.2	83.3
Personalization	0	0	0	0.85	0.87	0.87

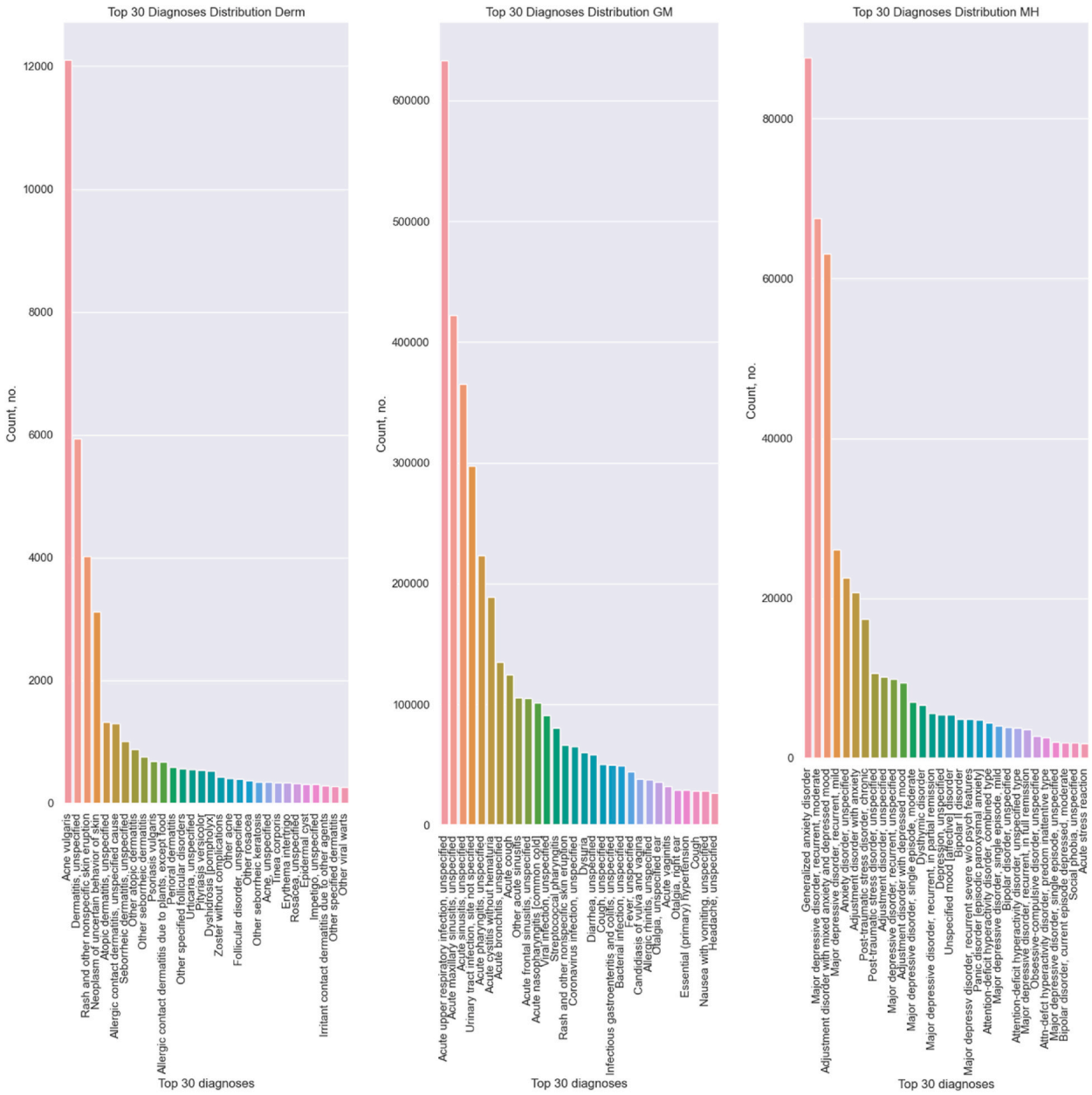


Fig. 2. Diagnosis distribution across clinical specialties Dermatology (left), General Medicine (center), Mental Health (right).

References

[1] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K.M. Abbas, R. Sundarsekar, Big data knowledge system in healthcare, Stud.Big Data 23 (2017) 133–157, https://doi.org/10.1007/978-3-319-49736-5_7/FIGURES/6.

[2] S.S. Lloyd, J.P. Rissing, Physician and coding errors in patient records, JAMA 254 (10) (1985) 1330–1336, https://doi.org/10.1001/JAMA.1985.03360100080018.

[3] M. Farzandipour, A. Sheikhtaheri, Accuracy of diagnostic coding based on ICD-10, KAUMS J.(FEYZ) 12 (4) (2009) 68–77. April 11, 2023, http://feyz.kaums.ac.ir/article-1-688-en.html.

[4] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, Measuring diagnoses: ICD code accuracy, Health Serv. Res. 40 (5p2) (2005) 1620–1639, https://doi.org/10.1111/J.1475-6773.2005.00444.X.

[5] E. Sezgin, S. Özkan, A systematic literature review on Health Recommender Systems, in: 2013 E-Health and Bioengineering Conference, EHB), 2013, pp. 1–4, https://doi.org/10.1109/EHB.2013.6707249.

[6] P.F. Chen, S.M. Wang, W.C. Liao, et al., Automatic ICD-10 coding and training system: deep neural network based on supervised learning, JMIR Med Inform 9 (8) (2021) e23230, https://doi.org/10.2196/23230.

[7] F. Teng, Z. Ma, J. Chen, M. Xiao, L. Huang, Automatic medical code assignment via deep learning approach for intelligent healthcare, IEEE J Biomed Health Inform 24 (9) (2020) 2506–2515, https://doi.org/10.1109/JBHI.2020.2996937.
[8] W. Ip, P. Prahalad, J. Palma, J.H. Chen, A data-driven algorithm to recommend initial clinical workup for outpatient specialty referral: algorithm development and validation using electronic health record data and expert surveys, JMIR Med Inform 10 (3) (2022) e30104, https://doi.org/10.2196/30104.
[9] Y. Bao, X. Jiang, An intelligent medicine recommender system framework, in: 2016 IEEE 11th Conference on Industrial Electronics and Applications, ICIEA), 2016, pp. 1383–1388, https://doi.org/10.1109/ICIEA.2016.7603801.
[10] J.G.D. Ochoa, O. Csiszár, T. Schimper, Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks, BMC Med Inform Decis Mak 21 (1) (2021) 186, https://doi.org/10.1186/s12911-021-01553-3.
[11] J.H. Chen, T. Podchiyska, R.B. Altman, OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records, J. Am. Med. Inf. Assoc. 23 (2) (2016) 339–348, https://doi.org/10.1093/jamia/ocv091.
[12] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang, K. Li, A disease diagnosis and treatment recommendation system based on big data mining and cloud computing, Inf. Sci. 435 (2018) 124–149, https://doi.org/10.1016/j.ins.2018.01.001.
[13] E. Moons, A. Khanna, A. Akkasi, M.F. Moens, A comparison of deep learning methods for ICD coding of clinical records, Appl. Sci. 10 (2020) 5262, https://doi.org/10.3390/APP10155262, 2020;10(15):5262.
[14] J.H.B. Kuo, C.C. Yeh, C.Y. Yang, et al., Applying deep learning model to predict diagnosis code of medical records, Diagnostics 13 (2023) 2297, https://doi.org/10.3390/DIAGNOSTICS13132297, 2023;13(13):2297.
[15] K. Goddard, A. Roudsari, J. Wyatt, Automation bias – a hidden issue for clinical decision support system use, Stud. Health Technol. Inf. 164 (2011) 17–22, https://doi.org/10.3233/978-1-60750-709-3-17.
[16] S. van der Walt, J. Millman, Python in science, in: S. van der Walt, J. Millman (Eds.), 9th Python in Science Conference, 2010, https://doi.org/10.3301/ROL.2016.64.
[17] Waskom M, Botvinnik O, Hobson P, et al. seaborn: v0.5.0 (November 2014). Published online November 14, 2014. doi:10.5281/ZENODO.12710.
[18] F. Pedregosa, M. Vincent, B. Thirion, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. March 10, 2019, http://scikit-learn.sourceforge.net.
[19] M. Abadi, A. Agarwal, P. Barham, et al., TensorFlow: large-scale machine learning on heterogeneous distributed systems, Published online March 14 (2016), https://doi.org/10.48550/arxiv.1603.04467.
[20] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, Ann. Intern. Med. 162 (1) (2015) 55–63, 10.7326/M14-0697/ASSET/IMAGES/LARGE/9FF3_FIGURE_3_TYPES_OF_PREDICTION_MODEL_STUDIES_COVERED_BY_THE_TRIPOD_STATEMENT.JPEG.
[21] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies, Int. J. Med. Inf. 153 (2021), https://doi.org/10.1016/j.ijmedinf.2021.104510.
[22] S.H. Noh, Analysis of gradient vanishing of RNNs and performance comparison, Information 12 (2021) 442, https://doi.org/10.3390/INFO12110442, 2021;12(11):442.
[23] A.K. Sahoo, C. Pradhan, R.K. Barik, H. Dubey, DeepReco: deep learning based health recommender system using collaborative filtering, Computation 7 (2019) 25, https://doi.org/10.3390/COMPUTATION7020025, 2019;7(2):25.
[24] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, R. Burke, Feedback loop and bias amplification in recommender systems, in: International Conference on Information and Knowledge Management, Proceedings, 2020, pp. 2145–2148, https://doi.org/10.1145/3340531.3412152, 4(20.
[25] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The connection between popularity bias, calibration, and fairness in recommendation, in: RecSys 2020 - 14th ACM Conference on Recommender Systems, 2020, pp. 726–731, https://doi.org/10.1145/3383313.3418487. Published online September 22.
[26] H. Abdollahpouri, Popularity bias in ranking and recommendation, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society vol. 27, 2019, pp. 529–530, https://doi.org/10.1145/3306618. Published online January.