A review of machine learning models for the prediction of sepsis in Canadian healthcare systems

Brett Caswell & Anna Luo
CPSC 340: Machine Learning and Data Mining
The Universty of British Columbia

Abstract

Sepsis is a life-threatening condition that is responsible for a significant number of deaths and healthcare costs in Canada. Early detection is crucial for improving patient outcomes, but diagnosing sepsis remains challenging due to its varied symptoms. Traditional diagnostic tools, such as SIRS, SOFA, and MEWS, have limitations in accuracy and timeliness. Recent advancements in artificial intelligence have shown promise in predicting sepsis earlier, using electronic health record data. This paper reviews several machine learning models, including Random Forest, XGBoost, and deep neural networks, highlighting their effectiveness in sepsis prediction. While these AI models outperform traditional methods in clinical trials, their current potential for implementation in Canada is limited. The review explores the challenges of adopting these technologies in Canadian healthcare, such as data diversity, interpretability, and the need to reduce false positives. We highlight strategies to integrate AI models into Canada's healthcare system to reduce the impact of sepsis on patient mortality and improve healthcare costs.

1 Introduction

Sepsis is a deadly and pervasive medical condition involved in one in every 18 deaths in Canada (Statistics Canada, 2016). Following an event such as infection or an injury, sepsis is caused by the body's unchecked release of immune mediators that lead to excessive inflammation, blood clotting, and leaks from blood vessels. The impact of the disease is ubiquitous throughout Canada, costing the healthcare system 1.7 billion dollars annually and representing a significant portion of hospital admissions (Sepsis Canada, 2024).

1.1 Diagnosis of sepsis

Rapid detection of sepsis enables opportune treatment of the condition, which can drastically improve health outcomes and costs to the healthcare system. However, the clinical manifestations of sepsis are diverse, making it an especially difficult condition to detect in a timely manner. To identify patients with sepsis, a myriad of scoring systems can be utilized. Systemic inflammatory response syndrome (SIRS) is one diagnostic criteria, which accounts for a patient's body temperature, heart rate, respiratory rate, and leukocyte count (Cleveland Clinic, 2024). The Sequential Organ Failure Assessment (SOFA) is a scoring system that assesses a patient's organ function through mean arterial pressure, glucose levels, bilirubin levels, PaO2/FiO2 ratio, platelet levels, and creatinine levels (Lambden et al., 2019). There also exists a faster version of SOFA known as quick SOFA (qSOFA) (Shahsavarinia et al., 2019). Similarly, the Modified Early Warning Score (MEWS) attempts to identify at-risk patients through use of blood pressure, heart rate, respiratory rate, body temperature and AVPU score (Hester et al., 2021). SIRS, SOFA, and MIRS are all criteria or scoring systems that are considered the gold standard in clinical diagnosis of sepsis.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1.2 Implementation of AI systems in sepsis prediction

Due to the importance of early detection of sepsis, a wealth of research has been conducted to develop artificial intelligence (AI) models capable of predicting the onset of sepsis using electronic health record (EHR) data (Islam et al., 2019). By leveraging AI to predict the emergence of sepsis in patients, clinicians can decrease the time required for treatment, thereby improving a patient's chance of recovery. Different algorithms and datasets have been applied to develop a useful AI model for the prediction of sepsis, and some models have been assessed in clinical settings to examine their ability to perform in real-world settings (Haas & McGill, 2022).

1.3 Objective

Despite the plethora of advancements in the field of artificial intelligence and sepsis prediction on an international scale, little research has been conducted to assess Canada's readiness and progress to adopt sepsis prediction models in healthcare settings. In this review, we aim to analyze the growing body of research on AI model development for sepsis prediction, summarize the impacts of existing work on the Canadian healthcare system, and suggest pathways forward that can help make Canada a leader in the development and adoption of this technology. Accordingly, we hope to inspire new research to close the gaps in the research landscape and reduce the countless number of lives lost to sepsis across the country.

2 Review

2.1 Development of machine learning models for sepsis prediction

2.1.1 Random Forest

Random Forest is an ensemble machine learning method that is a combination of random trees (Breiman, 2001). Each tree is trained on a bootstrapped sample of the training data, then for classification the random forest is formed together by majority voting across the collection of all trees. Due to the Law of Large Numbers overfitting does not occur if the model decides to combine more trees. The bootstrapping for the trees also allows us to maintain randomness in our model.

Giannini et al (2019) used a random forest classifier to predict severe sepsis and septic shock. The data used to train this random forest was non-ICU patient data obtained from three hospitals under the University of Pennsylvania Health System from July 2011 - June 2014. The model looked at 587 features, and was developed using 100 estimator trees and the gini index for splits. Conclusions from validation in a clinical setting at the same hospitals found a sensitivity of 0.26 and specificity of 0.98. The sensitivity measures the percentage of true positives out of all positively predicted outcomes, whereas sensitivity measures the percentage of true negatives out of all negatively predicted outcomes.

2.1.2 Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a supervised learning technique developed by Tianqi Chen that builds upon regular gradient boosting algorithms by adding regularization terms and other optimizing factors (Karale, 2024). Researchers have incorporated XGBoost techniques into their prediction algorithms and are finding promising results.

Lin et al (2021) developed a machine learning model using XGBoosting to identify sepsis patients in the emergency department. The model was trained and internally validated on emergency department data collected from Chi-Mei Medical Center in Taiwan. It was then externally validated on emergency department data from Taoyuan General Hospital in Taiwan. The test was done in a 80-20 training split, training using all features in the collected dataset. They also performed feature selection using XGBoost's built-in function in Python and found that the top 15 features accounted for explaining 65% of the total weight. The top 5 features were C-reactive protein (CRP), sodium (Na), percentage of lymphocytes in the differential of the complete blood count (Lympho), creatinine (Cr), and blood temperature (BT). The model performed well with an area under the receiver operating characteristic curve (AUROC) of 0.85 in internal validation and 0.75 in external validation.

2.1.3 Deep neural networks

Deep neural networks are often defined as the implementation of more than one layer of neurons in a neural network, organized in complex network architectures (Janiesch et al., 2021). This form of machine learning allows us to automatically discover learning representations that are appropriate for the model's underlying intended task using raw data, which demarcates it from other machine learning models. Deep neural networks are especially useful at making predictions with high-dimensional datasets, which can be especially helpful for the abundance of time-series data available in a patient's EMR (Janiesch et al., 2021).

Rafiei et al. (2021) developed a deep neural network for sepsis prediction that employs Long Short-Term Memory (LTSM), convolutional, and fully connected layers to predict the onset of sepsis. Using demographic data, vital signs, and the optional inclusion of laboratory tests from the 2019 PhysioNet/Computing in Cardiology Challenge dataset, the deep learning model achieved an AUROC of 0.89 and 0.86 for respectively predicting sepsis 4 and 12 hours before onset, using only vital signs and demographic variables. Additionally, the model achieved an AUROC of 0.92 and 0.84 for predicting sepsis 4 to 12 hours before onset when also supplied with laboratory test data (Rafiei et al., 2021).

Zargoush et al. (2021) also developed a deep learning model involving a bidirectional long short-term memory algorithm. Similarly to Rafiei et al. (2021), the study utilized publicly available data through the 2019 PhysioNet computing challenge. To provide a reference of the deep neural net's performance amongst other algorithms, six other types of models were created including logistic regression (LR), classification and regression trees (CART), XGBoost, naïve Bayes (NB), linear discriminant analysis (LDA), and AdaBoost (ADA). Many of the models performed close to the deep learning model in terms of accuracy and specificity, including the LDA, XGBoost, and LR models. However, the deep learning model outperforms all other machine learning models in sensitivity and AUROC values. The deep learning model resulted in a sensitivity and AUROC of 0.80 and 0.91, while no other model resulted in sensitivity values above 0.75 or an AUROC above 0.85.

2.2 Clinical implications

2.2.1 NAVOY Sepsis

NAVOY Sepsis is a machine learning algorithm that uses routinely collected vital parameters, blood gas values, and lab values as inputs to help identify patients in the ICU who are at a high risk of developing sepsis (Persson et al., 2024). It was deployed for a trial run in December 2020 to September 2021 at the ICU in the Skåne University Hospital Malmö where a total of 304 patients were studied. This NAVOY Sepsis algorithm was developed as a convolutional neural network using MIMIC-III data collected from the MIT Lab for Computational Physiology. The first convolutional layer contains 10 filters while the second has 5 each with size (1,2). There are then 4 fully connected layers with size 50, 25, 15, 10 respectively and one final output layer (Persson et al., 2021). The model resulted in an accuracy score of 0.79 (95% CI: 0.70, 0.88) , sensitivity of 0.80 (95% CI: 0.62, 0.98), specificity of 0.78 (95% CI: 0.68, 0.88), AUROC of 0.80, and positive predictive value of 0.53 (Persson et al., 2024).

In relation to similar studies of machine learning algorithms trained and used on ICU data, NAVOY Sepsis was found to have more accurate predictions than all of them. Algorithms such as the Sepsis Sniffer (Olenick et al., 2017), St.John Sepsis Surveillance Agent (Amland et al., 2017), and TREWscore (Henry et al., 2015) performed worse than NAVOY Sepsis in their positive predictive values, specificity value, and AUROC respectively.

2.2.2 InSight

Similarly to NAVOY Sepsis, Dascena, a company in the United States, has also built their own machine learning model called InSight. It is an early sepsis prediction tool that helps clinical workers identify patterns to predict the risk of a patient developing sepsis. They use continuous vital sign data from a patient and an alert is sent when their predicted risk of developing sepsis is above a certain threshold (Dascena, n.d.). InSight is different from many other algorithms in that it only takes in six core features – temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, oxygen saturation (Mao et al., 2018). It has a sensitivity and specificity of 0.90 each comparable to current traditional tools such as SIRS, SOFA, qSOFA who had sensitivities of 0.87, 0.70, and 0.54 respectively and specificity of 0.46, 0.78, and 0.89 respectively.

The success of InSight had gained the attention of the National Institute of Health, and in 2022 awarded Dascena a grant to continue to develop this algorithm into a new phase called HindSight (Leathers, 2022). This new algorithm was built to combat the problems of alert fatigue, reducing the number of false positives. The inner workings of the InSight algorithm has not been released to the public though they plan to develop the HindSight algorithm further by adapting it to the idiosyncrasies of real-world situations. It goes through periodic retraining allowing it to adapt to site-specific practices, and reduce the number of false positive alerts.

3 Discussion

3.1 Comparison of past work and potential improvements

The majority of the literature reviewed suggests that machine learning outperforms traditional severity scoring systems at predicting the onset of sepsis. This improved performance is likely due to the ability of machine learning algorithms to employ calculations outside of rule-based splitting that is often used by traditional scoring methods (Zargoush et al., 2021). While traditional scoring systems were developed primarily to assess the risk of sepsis, machine learning algorithms provide the enhanced capability of accurate sepsis prediction.

Evaluating the performance of different models, Lin et al. (2021) found that XGBoost attained a validation AUROC of 0.75, while traditional scoring systems and diagnostic criteria such as SIRS and qSOFA attained AUROC values of 0.57 and 0.66. However, our review suggests that deep neural networks can surpass the success of XGBoost and other "shallow" machine learning algorithms (Janeisch et al., 2021). This phenomenon is evidenced by the performance of the deep neural net model produced by Zargoush et al. (2021), which attained much better results than that of the other models considered in the study. Additionally, when summarizing the results of a deep learning algorithm against the results of the well-known gradient-tree boosting algorithm InSight, the model produced much more favourable AUROC scores, even when predicting more than four hours before its onset (Rafiei et al., 2021). While InSight has achieved improved sensitivity, specificity, and AUROC scores since its initial development through inventive strategies such as HindSight, the relative successes of deep learning models suggest that newer deep learning models should be subject to clinical validation and should explore the potential benefits of reinforcement learning (Dascena, n.d.; Leathers, 2022).

While the advancements of AI models in sepsis prediction are promising, there is much room for improvement. Healthcare practitioners hold many concerns over implementing emerging sepsis risk prediction tools, such as alert fatigue, clinical relevance, difficult explanations, confusing outputs, and expectation management (Joshi et al., 2022). As such, machine learning predictions models must improve the rate of false positives to reduce alert fatigue and tailor their development towards greater interpretability.

Many suggestions have been proposed to improve the utility of AI models for sepsis prediction. For instance, a recent paper by Gao et al. (2024) includes Shapley Additive Explanation analysis to their sepsis prediction model, which identifies and quantifies the contribution of each feature to the model's prediction, enhancing the ability of a clinician to interpret its results. We can also expand the use case of prediction models to incorporate suggestions for optimal treatment. A reinforcement learning model was developed to advise care providers on optimal treatment strategies, which underscores the potential for machine learning models to act as expert systems for the identification and treatment of sepsis (Komorowski et al., 2018).

3.2 Considerations for implemention in Canada

Many of these studies we have looked at so far have built models that are trained and validated using data split from the same sites or regions. The random forest trained and validated their model from the same three hospitals in Pennsylvania, but over the span of different years (Giannini et al., 2019). Similarly Lin et al (2021) trained and validated their extreme gradient boosting model on different sites, however, within the same jurisdiction.

Moor et al (2023) conducted a multi-centre cohort study looking at the datasets HiRID from Switzerland, AUMC the Netherlands and MIMIC-III from the United States. They created a deep learning-based early warning system, and trained their model for each dataset. They then internally validated with the held-out testing set to see how well they perform within their own cohorts. Secondly external validation was done using one cohort's model and validating it on the testing set of the other two cohort's separately. Averaging across the three cohorts, for internal validation the model obtained an AUC of 0.846 and for external validation obtained an AUC of 0.761. We can see that performance for external validation seems to drop from internal validation. However, if we added 10% of our data from the external site into our training algorithm, it was found to increase the AUC to 0.807 (Moor et al., 2023).

Looking at the results obtained by the multi-cohort study performed by Moor et al (2023), we can potentially use pre-existing deep learning models and train them using a portion of collected data in Canadian hospitals. This can allow us to still effectively predict sepsis in patients while not having to obtain vast amounts of data every time we want to implement these algorithms into a new environment. HindSight can also be evaluated to be clinically trialed in Canada, as it is an algorithm that will retrain itself with new data obtained (Leathers, 2022). The algorithm can be introduced in hospitals for a set amount of time without its predictions being utilized by medical staff, allowing itself to collect and retrain using new data at each site. This can allow the algorithm to learn the new behaviour in this setting and provide more accurate results later on.

However, it is also important to recognize that the data that is obtained for training models may underrepresent different groups of people. Canada is a multicultural country that has a publicly funded health care system, which allows people of all backgrounds and social standings to get healthcare. Many of these models we looked at were trained using the MIMIC-III dataset (Moor et al., 2023; Persson et al., 2024), collected in Massachusetts and the USA (Dascena, n.d.) in general. Since the USA has a private healthcare system these dataset may disproportionately represent different populations of people. The increase in the use of electronic health records allow for implementation of machine learning algorithms to be implemented within our systems. In Canada nearly all clinicians use a form of electronic medical records, and there has been immense growth these past few years. In 2022, it was reported that 93% of primary care physicians used electronic medical records compared to 73% in 2015 (CIHI, 2024). Electronic medical records are essential in building models for these machine learning algorithms, allowing researchers to have access to the data in an efficient and clear manner. Additionally hospitals that have yet to move to electronic medical records may fall behind as a lack of a digitized system may hinder their ability to supply data that these supervised machine learning algorithms rely heavily on for training.

3.3 Conclusion

This study conducted a literature review of AI models used for the prediction of the onset of sepsis and addresses the obstacles of applying previous work to the Canadian healthcare system. First, we examined several existing machine learning models and analyzed their performance from both statistical and clinical perspectives. Then, we further analyzed the relative performance of various types of AI models in predicting sepsis, noting their strengths, limitations, and obstacles to widespread clinical adoption. Finally, we elucidated the existing barriers to adopting AI models for sepsis prediction specifically for the Canadian healthcare setting. By highlighting the progress and shortcomings of previous research and communicating paths forward to clinical adoption of AI sepsis models in Canada, this review contributes to the growing body of literature in hopes of reducing the countless lives and public funds lost to sepsis in Canada.

References

[1] Amland, R. C.,& Sutariya, B. B. (2017). Quick sequential [sepsis-related] organ failure assessment (qSOFA) and St. John Sepsis Surveillance Agent to detect patients at risk of sepsis: An observational cohort study. *American Journal of Medical Quality, 33*(1), 50–57. https://doi.org/10.1177/1062860617692034

[2] Breiman, L. (2001). Random Forests. (R. E. Schapire, Ed.). *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

[3] *Canadians and health care providers want connected Electronic Health Information Systems*. CIHI. (2024). https://www.cihi.ca/en/taking-the-pulse-measuring-shared-priorities-for-canadian-health-care-2024/canadians-and-health-care-providers-want-connected-electronic-health-information

[4] Dascena, (n.d.) https://f.hubspotusercontent30.net/hubfs/4120680/InSight_Fact_Sheet.pdf

[5] Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., & Pishgar, M. (2024). Prediction of sepsis mortality in ICU patients using Machine Learning Methods. *BMC Medical Informatics and Decision Making, 24*(1). https://doi.org/10.1186/s12911-024-02630-z

[6] Giannini, H. M., Ginestra, J. C., Chivers, C., Draugelis, M., Hanish, A., Schweickert, W. D., Fuchs, B. D., Meadows, L., Lynch, M., Donnelly, P. J., Pavan, K., Fishman, N. O., Hanson, C. W., & Umscheid, C. A. (2019). A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice*. *Critical Care Medicine, 47*(11), 1485–1492. https://doi.org/10.1097/ccm.0000000000003891

[7] Haas, R., & McGill, S. C. (2022). *Artificial Intelligence for the prediction of sepsis in adults.* National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK596676/

[8] Henry, K. E., Hager, D. N., Pronovost, P. J., & Saria, S. (2015). A targeted real-time early warning score (TREWScore) for Septic Shock. *Science Translational Medicine, 7*(299). https://doi.org/10.1126/scitranslmed.aab3719

[9] Hester, J., Youn, T. S., Trifilio, E., Robinson, C. P., Babi, M. A., Ameli, P., Roth, W., Gatica, S., Pizzi, M. A., Gennaro, A., Crescioni, C., Maciel, C. B., & Busl, K. M. (2021). *The Modified Early Warning Score: A Useful Marker of Neurological Worsening but Unreliable Predictor of Sepsis in the Neurocritically Ill-A Retrospective Cohort Study*. Critical care explorations, 3(5), e0386. https://doi.org/10.1097/CCE.0000000000000386

[10] *Home*. Sepsis Canada. (n.d.). https://www.sepsiscanada.ca/: :textĀdditionally%2C%20it%20is%20estimated %20that%20sepsis %20costs%20Canadians%20%241.7B%20annually

[11] Islam, M. M., Nasrin, T., Walther, B. A., Wu, C. C., Yang, H. C., & Li, Y. C. (2019). Prediction of sepsis patients using machine learning approach: A meta-analysis. *Computer methods and programs in biomedicine*, 170, 1–9. https://doi.org/10.1016/j.cmpb.2018.12.027

[12] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

[13] Joshi, M., Mecklai, K., Rozenblum, R., & Samal, L. (2022). Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. *JAMIA open, 5*(2), ooac022. https://doi.org/10.1093/jamiaopen/ooac022

[14] Karale, J. (2024, April 22). *Understanding the difference between GBM vs XGBoost*. The Talent500 Blog. https://talent500.com/blog/understanding-the-difference-between-gbm-vs-xgboost/

[15] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine, 24*(11), 1716–1720. https://doi.org/10.1038/s41591-018-0213-5

[16] Lambden, S., Laterre, P. F., Levy, M. M., & Francois, B. (2019). The SOFA score-development, utility and challenges of accurate assessment in clinical trials. *Critical care (London, England), 23*(1), 374. https://doi.org/10.1186/s13054-019-2663-7

[17] Leathers, K. (2022, March 29). *Dascena awarded NIH grant to better predict the onset of sepsis in patients*. Business Wire. https://www.businesswire.com/news/home/20220329005048/en/Dascena-Awarded-NIH-Grant-to-Better-Predict-the-Onset-of-Sepsis-in-Patients

[18] Lin, P.-C., Chen, K.-T., Chen, H.-C., Islam, Md. M., & Lin, M.-C. (2021). Machine learning model to identify sepsis patients in the emergency department: Algorithm Development and Validation. *Journal of Personalized Medicine*, 11(11), 1055. https://doi.org/10.3390/jpm11111055

[19] Mao, Q., Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., Shieh, L., Chettipally, U., Fletcher, G., Kerem, Y., Zhou, Y., & Das, R. (2018). Multicentre validation of a sepsis prediction algo-

rithm using only vital sign data in the Emergency Department, General Ward and ICU. *BMJ Open*, 8(1). https://doi.org/10.1136/bmjopen-2017-017833

[20] Moor, M., Bennett, N., Plečko, D., Horn, M., Rieck, B., Meinshausen, N., Bühlmann, P., & Borgwardt, K. (2023). Predicting sepsis using deep learning across international sites: A retrospective development and validation study. *eClinicalMedicine*, 62, 102124. https://doi.org/10.1016/j.eclinm.2023.102124

[21] Olenick, E. M., Zimbro, K. S., D'Lima, G. M., Ver Schneider, P., & Jones, D. (2017). Predicting sepsis risk using the "Sniffer" algorithm in the Electronic Medical Record.*Journal of Nursing Care Quality, 32*(1), 25–31. https://doi.org/10.1097/ncq.0000000000000198

[22] Persson, I., Macura, A., Becedas, D., & Sjövall, F. (2024). Early prediction of sepsis in intensive care patients using the machine learning algorithm NAVOY® sepsis, a prospective randomized clinical validation study. *Journal of Critical Care, 80*, 154400. https://doi.org/10.1016/j.jcrc.2023.154400

[23] Persson, I., Östling, A., Arlbrandt, M., Söderberg, J., & Becedas, D. (2021). A machine learning sepsis prediction algorithm for intended intensive care unit use (NAVOY sepsis): Proof-of-concept study. *JMIR Formative Research, 5*(9). https://doi.org/10.2196/28000

[24] Rafiei, A., Rezaee, A., Hajati, F., Gheisari, S., & Golzan, M. (2021). SSP: Early prediction of sepsis using fully connected LSTM-CNN Model. *Computers in Biology and Medicine*, 128, 104110. https://doi.org/10.1016/j.compbiomed.2020.104110

[25] *SIRS (Systemic Inflammatory Response Syndrome)*. Cleveland Clinic. (2024, May 1). https://my.clevelandclinic.org/health/diseases/25132-sirs-systemic-inflammatory-response-syndrome

[26] Shahsavarinia, K., Moharramzadeh, P., Arvanagi, R. J., & Mahmoodpoor, A. (2020). qSOFA score for prediction of sepsis outcome in emergency department. *Pakistan journal of medical sciences*, 36(4), 668–672. https://doi.org/10.12669/pjms.36.4.2031

[27] Statistics Canada. (2016, November 23). *Health at a glance. Deaths involving sepsis in Canada.* https://www150.statcan.gc.ca/n1/pub/82-624-x/2016001/article/14308-eng.htm

[28] Zargoush, M., Sameh, A., Javadi, M., Shabani, S., Ghazalbash, S., & Perri, D. (2021). The impact of recency and adequacy of historical information on sepsis predictions using machine learning. *Scientific Reports, 11*(1). https://doi.org/10.1038/s41598-021-00220-x