Peer Review File

Article information: https://dx.doi.org/10.21037/tp-23-25

Reviewer A

The authors from the National University of Singapore and KK Women's and Children's Hospital present a scoping review of the literature on machine learning and probabilistic graphical modeling for pediatric sepsis. The manuscript covers the currently literature on this topic quite comprehensively, with an extensive list of relevant articles. However, there are several areas that would need improvement prior to publication of this manuscript.

General Recommendations:

Comment 1. The intention of the paper is to compare ML to PGM methodology, but the data are not presented in a way that makes it easy for the reader to see this comparison.

Reply: Thank you for your comment. In the revised manuscript, we present the comparison in two approaches: (1) an overall comparison between PGM and other ML methods, and (2) a deeper comparison based on studies using both PGM and other ML methods on the same dataset.

For the first approach, we added an additional table (**Table 4**) to describe the advantages and disadvantages between different ML methods. For PGM, we included common methods described in the literature: Tree Augmented Naïve Bayes (TAN), Naïve Bayes (NB), Hidden Markov Model (HMM), Influence Diagram (ID), Bayesian Network (BN), and Dynamic Bayesian Network (DBN). For other ML methods, we included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) as these are the most commonly used methods in our review. We observed that all methods exhibit different strengths in different areas. The common drawback of all methods is that they are sensitive to outliers, imbalance data, and are easily prone to overfitting when the model settings are not accurately figured. In addition, there are trade-offs between model complexity, training time, and interpretability. Models with a high level of complexity may require a long training period and become challenging to explain.

In our second approach, we compared PGM and other ML methods on the same dataset. Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics have been reported across the studies. Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results.

Characteristic comparison between PGM methods and other ML methods.

Characteristics	PGM Methods						Other ML Methods					
	TAN	NB	HMM	ID	BN	DBN	LR	RF	SVM	NN	XGBoost	DT
Data Handling												
Handling small data size	*	*	*	*	*	*			*			
Handling big dataset	*	*	*	*	*	*	*			*		
Handling missing data	*	*	*	*	*	*		*	*		*	*
Handling imbalance data												
Handling noisy data	*	*	*	*	*	*			*	*		
Handling outliers								*	*		*	*
Usage on continuous data									*	*	*	
Usage on category data	*	*	*	*	*	*	*	*				*
Usage on time-series data			*			*			*	*		
Variable selection								*	*	*	*	*
Presentation												
Visualization	*	*	*	*	*	*		*			*	*
Capability												
Classification	*	*	*		*	*	*	*	*	*	*	*
Regression								*	*	*	*	*
Causal inference	*	*	*	*	*	*						
Support decision-making				*				*			*	*

Natural language processing	*	*	*		*	*	*	*	*	*		*
Image processing	*	*	*		*	*	*	*	*	*		*
Interpretation												
Explainable method	*	*	*	*	*	*	*	*			*	*
Computational requirement												
Require hardware dependency									*	*	*	
Require more training time								*	*	*		
Prone to overfitting	*	*	*	*	*	*	*	*	*	*	*	*

It is important to note that each method exhibits different strengths in different areas. In this table, the characteristics described are not meant to be exhaustive. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them.

Abbreviation: **BN** Bayesian Network, **DBN** Dynamic Bayesian Network, **DT** Decision Tree, **HMM** Hidden Markov Model, **ID** Influence Diagram, **LR** Logistic Regression, **ML** machine learning, **NB** Naïve Bayes, **NN** Neural Network, **PGM** probabilistic graphical model, **RF** Random Forest**, SVM** Support Vector Machine, **TAN** Tree Augmented Naïve Bayes, **XGBoost** Extreme Gradient Boosting.

Performance comparison of PGM and other MLs on the same dataset

Publication	Methods used	AUC	SEN	SPE	NPV	PPV
Mani et al., 2014	RF	0.57-0.65	0.82-0.94	0.18-0.47	0.28-0.73	0.55-0.70
	SVM	0.61-0.68	0.79-0.88	0.18-0.26	0.27-0.59	0.51-0.69
	KNN	0.54-0.62	0.83-0.86	0.18-0.29	0.30-0.55	0.52-0.70
	CART	0.65-0.77	0.75-0.81	0.18-0.30	0.23-0.51	0.51-0.68
	LR	0.61	0.86-0.87	0.18-0.33	0.35-0.57	0.52-0.72

	LBR	0.58-0.62	0.86-0.85	0.18-0.33	0.36-0.52	0.52-0.72
	AODE	0.53-0.61	0.85-0.88	0.18-0.36	0.38-0.54	0.52-0.73
	NB	**0.64-0.78**	**0.83-0.95**	**0.18-0.47**	**0.31-0.76**	**0.55-0.72**
	TAN	**0.53-0.59**	**0.84**	**0.18-0.32**	**0.32-0.52**	**0.50-0.72**
Gomez et al., 2019	Adaboost	0.943	0.944	0.944	0.942	0.945
	Bagged Trees	0.88	0.901	0.858	0.896	0.866
	RF	0.84	0.861	0.818	0.853	0.827
	LR	0.787	0.771	0.804	0.777	0.8
	SVM	0.755	0.641	0.868	0.705	0.831
	DT	0.751	0.816	0.687	0.788	0.726
	KNN	0.64	0.565	0.715	0.62	0.667
	NB	**0.666**	**0.431**	**0.901**	**0.61**	**0.814**
Masino et al., 2019	Adaboost	0.83-0.85	0.8	0.72	0.92	0.51
	GB	0.8-0.87	0.8	0.74	0.92	0.53
	GP	0.75-0.79	0.8	0.6	0.9	0.44
	KNN	0.73-0.79	0.8	0.55	0.9	0.39
	LR	0.83-0.85	0.8	0.74	0.93	0.52
	RF	0.82-0.86	0.8	0.74	0.92	0.53
	SVM	0.82-0.86	0.8	0.72	0.92	0.51
	NB	**0.81-0.84**	**0.8**	**0.73**	**0.92**	**0.52**
Song et al., 2020	LR	0.86	-	-	0.94-0.96	0.4-0.5
	DT	0.6-0.84	-	-	0.84-0.95	0.39-0.57
	AdaBoost	0.81-0.83	-	-	0.91-0.94	0.41-0.53
	ET	0.80	-	-	0.81-0.88	0.53-0.68
	Bagging	0.77-0.81	-	-	0.83-0.88	0.45-0.59
	RF	0.81-0.82	-	-	0.83-0.88	0.51-0.66

	GB	0.86-0.87	-	-	0.92-0.94	0.45-0.56
	GNB	**0.81-0.82**	**-**	**-**	**0.95-0.96**	**0.28-0.38**
Cabrera-Quiros et al., 2021	LR	0.79	0.78	0.8	-	0.82
	NN	0.7	0.67	0.74	-	0.73
	NB	**0.71**	**0.68**	**0.74**	**-**	**0.73**

In this table, PGM performance is highlighted in bold. Metrics presented in the table (AUC, SEN, SPE, NPV, PPV) were chosen as they were reported in the respective studies. We excluded Stanculescu et al. (2014), Honore et al. (2020), Ying et al. (2021), and Kausch et al. (2021) from this table because they only reported AUC.

Abbreviations: AODE Averaged one dependence estimators, AUC area under the curve, CART Classification and Regression Tree, DT Decision Tree, ET Extra Trees, GB Gradient Boosting, GP Gaussian Process, KNN K-Nearest Neighbour, LBR Lazy Bayesian rules, LR Logistic Regression, ML Machine Learning, NB Naïve Bayes, NN Neural Network, NPV negative predicted value, PGM Probabilistic Graphical Model, PPV positive predicted value, RF Random Forest, SEN sensitivity, SPE specificity, SVM Support Vector Machine, TAN Tree Augmented Naïve Bayes.

Changes in text: Several changes have been made in the manuscript:

In the Abstract Result section: "**When applied to the same dataset, the performance of probabilistic graphical models (area under curve: 0.53-0.84, sensitivity: 0.43-0.95, specificity: 0.18-0.90, negative predictive value: 0.31-0.96, positive predictive value: 0.28-0.81) overlapped with other machine learning models (area under curve: 0.54-0.94, sensitivity: 0.57-0.94, specificity: 0.18-0.94, negative predictive value: 0.23-0.96, positive predictive value: 0.39-0.95)**".

In the Method section, page 8, line 142-147: "**In this review, we present the performance comparison in two approaches: (1) an overall qualitative comparison between PGM and other ML methods, and (2) a deeper comparison based on studies using both PGM and other ML methods on the same dataset. Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics have been reported commonly across several studies. In addition to the performance comparison, an analysis of pediatric sepsis definitions was also conducted from selected publications."**

In the Result section, page 9, line 179-186: "**Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results. Furthermore, the studies that used both PGM and other ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE), whereas other attributes, such as explainability and visualization, were not examined."**

The following paragraph has been added to the Discussion page 13, line 263-281: "**A comparison of the characteristics of PGM and other ML methods is presented in Table 4. Among the popular methods that are often used in the literature for PGM, we selected BN, NB, TAN, HMM, DBN, and Influence Diagram (ID). Other ML methods that we chose from our review include LR, RF, Support Vector Machine (SVM), NN, Extreme Gradient Boosting (XGBoost), and DT. We observed that all methods exhibit different strengths in different areas. For instance, several of the other ML methods shown in Table 4 are capable of performing both classification and regression, while PGM can only perform classification. The NN and the SVM excel on several criteria; however, they require longer training time, hardware dependence, and additional aids for visualization and interpretation. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them. The main disadvantage of all methods is that they are sensitive to outliers, imbalanced data, and easily prone to overfitting when the model settings are not probably figured. The complexity of the model, the training time, and the interpretability of the results are also subjected to trade-offs. It is likely that highly complex models will require a larger amount of training time and be more challenging to interpret**."

Comment 2. The level of detail provided is inconsistent. Some concepts that are rather basic are described in significant detail, while more complex concepts that may be less familiar to readers, are glossed over. To improve readability, be sure to clearly define / describe complex concepts or methodologies. For example, there is a lengthy discussion of sensitivity, specificity etc., but relatively little on different ML methodologies.

Reply: In response to your recommendation, we have summarized the evaluation metrics and added a brief description of different machine learning approaches.

Changes in text: The following paragraph have been added to Page 10-11, Line 196-207: "**Our findings have led us to make the following observations about ML approaches to pediatric sepsis. The first observation is that LR, RF, and NN were the most common ML methods, followed by the tree-based models. Similar to Linear Regression (fitting of a regression line to the data), the LR concept uses the sigmoid function in order to fit an S-curve to the data. This S-curve will determine the probability of the outcome, taking into account other variables, and whether it will pass the decision threshold. The RF consists of several small decision trees that work together as an ensemble, where the final decision is determined by a majority vote. Finally, a neural network is a structure of interconnected nodes nested in several layers, where each node is associated with a weight and an activation function. The data can only pass through the node when it is activated. Although most neural**

networks are feed-forward, meaning information can only be transmitted in one direction, feed-forward back-propagation networks are also available for adjusting the weights of the system".

Comment 3. There was a lot of speculation and broad recommendations that weren't necessarily based in the data that were presented. Try to keep the commentary to concepts that tie directly to the data.

Reply: Thank you for your comment. We have reviewed our manuscript and reduced broad recommendation. Where it is applicable, we have added the evidence from our findings to support our arguments.

Changes in text: Changes have been made in several places throughout the manuscript. These changes have been marked as "bold" with red color in the revised manuscript.

Specific Recommendations:

Comment 1. In the abstract results section, the final sentence is opinion of the authors, not true results. This should be removed and included in the discussion section of the manuscript.

Reply: We have amended the final sentence of the Abstract results section to reflect the performance of the probabilistic graphical model in comparison to other machine learning models.

Changes in text: The sentence was amended in the Abstract results section at Page 3, Line 49-54 as follows: "**When applied to the same dataset, the performance of probabilistic graphical models (area under curve: 0.53-0.84, sensitivity: 0.43-0.95, specificity: 0.18-0.90, negative predictive value: 0.31-0.96, positive predictive value: 0.28-0.81) overlapped with other machine learning models (area under curve: 0.54-0.94, sensitivity: 0.57-0.94, specificity: 0.18-0.94, negative predictive value: 0.23-0.96, positive predictive value: 0.39-0.95).**".

Comment 2. In the "Highlight Box" – the 2nd bullet point in the What is known and what is new section states that the review demonstrates why PGMs have redeeming qualities. However, this is not clear from the data presented and instead describes general qualities of PGMs not specific to the studies identified.

Reply: Thank you for your comment. The selected studies that used both PGM and ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE, NPV, PPV), whereas other attributes, such as explainability and visualization, were

not examined. Hence, we could not extract this information from the studies. Instead, we have highlighted this point in the result section and amended the sentences in the Highlight Box.

Changes in text: Several changes have been made to the manuscript:

In the Highlight box, "**PGMs exhibit certain qualities that distinguish them from other machine learning methods. These include interactive representation, transparent reasoning, missing data handling and a variety of methods, which can be utilized for various pediatric sepsis applications. There is, however, a lack of in-depth discussion of these aspects in comparison to other ML methods in the current literature".**

The following sentence has been added to the Result section, page 9, line 184-186: "**The studies that used both PGM and ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE, NPV, PPV), whereas other attributes, such as explainability and visualization, were not examined**".

Comment 3. It is not necessary to describe how the discussion section is organized – consider removing lines 191-196 (page 8).
Reply: Thank you for your suggestion. We have removed the respective lines.
Changes in text: The following sentences have been removed from the manuscript: "**Here we present some qualities of PGM that make it a noteworthy candidate for future studies. Then we summarize the usage of sepsis definitions in these studies and argue why a combined definition might improve sepsis prediction. Finally, we discuss the limitations of our review and provide directions for future studies.**"

Comment 4. Findings highlighted in the discussion are mixed in with recommendations. Consider restructuring so that there is a "recommendations" section that summarizes the salient recommendations.

Reply: We have moved the sentences with findings to the results section. We also added a subsection "Recommendation" in the Discussion as suggested.

Changes in text: The following sentences have been moved to the Results section, Page 8, Line 159: "**We observed a surge in Sepsis-3 use (n = 1 before 2020, n = 5 after 2020), even though positive cultures continue to be the most widely used definition (n = 8 before 2020, n = 8 after 2020).**"

The following sentences have been added to the new subsection "Recommendation" in the discussion section.
"In view of the PGM's inherent qualities and potential for clinical use, we recommend that researchers consider using PGM methods in future studies of pediatric sepsis. Additionally, we recommend that ML studies include more than one definition of sepsis in order to enhance their predictive capabilities. In the event that a dataset does not

contain enough information to extract more than one definition, researchers may consider combining several datasets. More granular data should also be collected in future original studies on sepsis to facilitate the extraction of multiple sepsis definitions. Finally, it would be desirable to conduct studies that could compare single definition-learned model with the multiple definitions-learned model in order to validate our hypothesis that combining two or more sepsis definitions will improve performance of ML methods."

Comment 5. On page 10, lines 257-259, I'm not sure what this statement means – what are the "similar arguments" and why is PGM superior in these other diseases? Please clarify this.
Reply: We observed the same situation in other diseases, such as adult sepsis, cancer, or cardiovascular that NB performance is inferior to its PGM relative methods, such as Bayesian Network, Dynamic Bayesian Network or Markov Models. By 'similar argument', we meant that NB might not be the best choice to model complex data where the association between variables are highly correlated. We have rewritten the paragraph to make our argument clearer.
Changes in text: The following amendment has been made in Page 13, Line 260-262: "**We observed the same situation in other diseases (e.g., adult sepsis, cancer, or cardiovascular) that NB performance is inferior to its PGM relative methods, such as Bayesian Network (BN), Dynamic Bayesian Network (DBN) or Hidden Markov Models (HMM).".**

Comment 6. Much of discussion reads like a general review article about PGM (pages 10-11, lines 264-297). Consider focusing on more on how the strengths / weaknesses apply specifically to the data and articles described. Alternatively, consider writing a review article about the different approaches to sepsis diagnoses using ML.
Reply: Thank you for your suggestion. We will consider your suggestion to write another review article that takes a deeper dive into the ML-based approaches to sepsis diagnosis as our next manuscript. In this revised manuscript, we added an additional table (Table 4) to describe the advantages and disadvantages between different ML methods. For PGM, we included common methods described in the literature: Tree Augmented Naïve Bayes (TAN), Naïve Bayes (NB), Hidden Markov Model (HMM), Influence Diagram (ID), Bayesian Network (BN), and Dynamic Bayesian Network (DBN). For other ML methods, we included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) as these are the most commonly used methods in our review. We observe that all methods exhibit different strengths in different areas. The common drawback of all methods is that they are sensitive to outliers, imbalance data, and are easily prone to overfitting when the model settings are not accurately figured. In addition, there are trade-offs between model complexity, training time, and interpretability. Models with a high level of complexity may require a long training period and become challenging to explain.

Changes in text: Several changes have been made in the manuscript:

We have added **Table 4** to compare the different characteristics of PGM and other ML methods.

In the Result section, page 9, line 179-186: "**Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different**

settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results. Furthermore, the studies that used both PGM and other ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE), whereas other attributes, such as explainability and visualization, were not examined."

The following paragraph has been added to the Discussion page 13, line 263-281: "**A comparison of the characteristics of PGM and other ML methods is presented in Table 4. Among the popular methods that are often used in the literature for PGM, we selected BN, NB, TAN, HMM, DBN, and Influence Diagram (ID). Other ML methods that we chose from our review include LR, RF, Support Vector Machine (SVM), NN, Extreme Gradient Boosting (XGBoost), and DT. We observe that all methods exhibit different strengths in different areas. For instance, several of the other ML methods shown in Table 4 are capable of performing both classification and regression. The NN and the SVM excel on several criteria; however, they require longer training time, are hardware-dependent, and require additional aids for visualization and interpretation. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and window-based data to use with methods that do not natively support them. The main disadvantage of all methods is that they are sensitive to outliers, imbalanced data, and easily prone to overfitting when the model settings are not accurately figured. The complexity of the model, the training time, and the interpretability of the results are also subjected to tradeoffs. It is likely that highly complex models will require a larger amount of training time and be more challenging to interpret."**

Comment 7. Page 12, line 306 – the surviving sepsis campaign was launched in 2002, with multiple iterations over the years since then.
Reply: The criteria of organ dysfunction have been applied to adult population first in 2016 and later to pediatric population in 2020. From the literature search, we observed a surge in using sepsis-3 definition after 2020 (n = 1 before 2020, n = 5 after 2020). Hence, we wanted to emphasize the recognition of dysfunctional organ criteria in diagnosing pediatric sepsis.
Changes in text: No changes in text.

Comment 8. Although I agree wholeheartedly that the lack of a "gold standard" definition for pediatric sepsis is a significant challenge, I have major concerns about the suggestion that combining multiple definitions is a solution to this. I worry this would only add confusion, further dilute the cohort being studied (specifically by adding more false positive cases) and decrease reproducibility.

Reply: We understand the reviewer's concern that this would add confusion and decrease the reproducibility. However, as the pediatric sepsis cases consist of several subgroups and their characteristics overlapped in certain ways, we think that future studies should consider this clinical pattern as well. As for ML models, given that no single definition is capable of identifying sepsis effectively, a model trained on a singular definition may not be sufficiently generalizable. We hypothesize that with a proper structure and set up, the machine learning methods may have the capability to learn the characteristics of these subgroups. However, this approach will need to be tested and validated by future studies. We have elaborated in our manuscript to address the reviewer's concern.

Changes in text: The following sentences have been added to Page 16, Line 348-350: "**We hypothesize that, with a proper structure and set up, ML methods may have the capability to learn the characteristics of these subgroups. However, this approach needs to be tested and validated in future studies.**".

Comment 9. In the conclusions section, please limit author opinion and focus on what was found in the data/significant of the findings.
Reply: Thank you for your comment. We have revised the conclusion section to be more objective.

Changes in text: The conclusion section has been amended on Page 18, Line 379-382: "**This scoping review summarizes ML and PGM approaches in pediatric sepsis over the past two decades and discusses how sepsis definitions were applied in these studies. The performance of PGM was comparable to other ML methods and can be considered as a potential tool for future pediatric sepsis studies and application.**".

Comment 10. Figure 2(c) – the final component of the model is repeated.
Reply: We have removed the repeated component in the model of Figure 2(c).
Changes in text: Figure 2(c) has been updated.

Comment 11. Figure 3 – I think the intention of this is to demonstrate how the 3 different definitions are similar and different, but the Venn diagrams are too busy with too small font for this to be effective.
Reply: Yes, Figure 3 was indeed to demonstrate the differences in 3 definitions of pediatric sepsis. We have adjusted the font size of the figures so that it can be viewed clearer.
Changes in text: The font size of Figure 3 has been updated.

Comment 12. Table 3 – Bacteremia = bacteria identified in the blood stream. You cannot diagnose a patient with bacteremia without a positive culture.
Reply: Thank you for your comment. We have corrected the definition for bacteremia, which is blood stream infections with positive cultures.

Changes in text: The correction was reflected in Table 2 as follows.

Table 2: Pediatric sepsis definition

Definitions		**Description**
IPSCC (2005)		SIRS and presence of suspected or proven infections
Positive cultures		Positive cultures of blood, CSF, etc.
Adapted Sepsis-3 (2016)		Dysregulated host response to infection and dysfunctional organs measured by age-based pSOFA
General Infections	Bacteremia	Blood stream infections with positive cultures
	Bacterial Infection	Bacterial infection with or without positive cultures
	Viral Infection	Viral infection with or without positive cultures
	Fungal Infection	Invasive fungal infection with or without positive cultures

Abbreviations: **CSF** Cerebrospinal fluid, **IPSCC** International pediatric sepsis consensus conference, **SIRS** Systemic Inflammatory Response Syndrome, **pSOFA** pediatric Sequential organ failure assessment.

Reviewer B

I have read your manuscript with interest. Overall, it is well written. However, few important issues need to be addressed.

Comment 1. Although the article seems to focus on improving diagnosis of sepsis, several of the articles chosen focus on use of ML to predict sepsis progression or risk. I would clarify your question and choose only those articles relevant.

Reply: Following your comment, we have reviewed the manuscript and amend the aim of our literature review. It was our intention to identify the state of the art of the probabilistic graphical model (PGM) in comparison to other machine learning (ML) approaches in pediatric sepsis application. For this reason, we have amended the objective sentence of our literature review.

Changes in text:

The objective sentence in the Abstract has been amended to: "**This scoping review aims to evaluate the feasibility of probabilistic graphical model methods in comparison to other machine learning approaches in pediatric sepsis application".**

The following changes has been made to Page 6, Line 98-99: "**Thus, in this scoping review, we aim to (1) evaluate the feasibility of PGM in pediatric sepsis application and (2) describe how pediatric sepsis definitions are used in the ML literature**".

Comment 2. Classification and Regression Tree analysis is perhaps the most used ML tool in the previous decade, used extensively by the late Dr. Hector Wong and colleagues for risk stratification of sepsis patients. You have included studies on risk stratification among septic patients and those that have included XGBoost. If you address point 1 and choose to include sepsis risk stratification, including all studies that utilize CART or TreeNet in addition to LR, RF, etc. is essential.

Reply: Thank you for your comment. We have amended our objective sentences throughout the manuscript and have added the publications related to sepsis risk stratification in our scoping review. We have found several studies conducted by Dr. Hector Wong et al. and other studies that utilized CART that were used to construct the risk model.
Changes in text: Several changes have been made throughout the manuscript to adjust the number of included studies. The following sentences was added to the Result section, page 9, line 171-174: "**In addition, other tree-based models, including Classification and Regression Tree (CART, n=6), Decision Tree (DT, n=3), Gradient Boosted Decision Tree (GBDT, n=8), Extra trees (ET, n=3), Bagged Trees (n=1), are also frequently used".**

Comment 3. How was the conclusion that PGM models were inferior to other ML models. The AUROCs show significant overlap. Did you use Net Reclassification or DeLong's test to compare AUROCs. The difference in number of studies using PGM vs. ML models are likely to bias against PGM models.
Reply: We did not have enough information to run the Net Reclassification or DeLong's test to compare the AUROCs. Therefore, we have presented the comparison in two approaches: (1) an overall comparison between PGM and other ML methods, and (2) a deeper comparison based on studies using both PGM and other ML methods on the same dataset.

For the first approach, we added an additional table (**Table 4**) to describe the advantages and disadvantages between different ML methods. For PGM, we included common methods described in the literature: Tree Augmented Naïve Bayes (TAN), Naïve Bayes (NB), Hidden Markov Model (HMM), Influence Diagram (ID), Bayesian Network (BN), and Dynamic Bayesian Network (DBN). For other ML methods, we included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) as these are the most commonly used methods in our review. We observed that all methods exhibit different strengths in different areas. The common drawback of all methods is that they are sensitive to outliers, imbalance data, and are easily prone to overfitting when the model settings are not accurately figured. In addition, there are trade-offs between model complexity, training time, and interpretability. Models with a high level of complexity may require a long training period and become challenging to explain.

In our second approach, we compared PGM and other ML methods on the same dataset. Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics have been reported across the studies. Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable

performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results.

Characteristic comparison between PGM methods and other ML methods.

Characteristics	**PGM Methods**						**Other ML Methods**					
	TAN	NB	HMM	ID	BN	DBN	LR	RF	SVM	NN	XGBoost	DT
Data Handling												
Handling small data size	*	*	*	*	*	*			*			
Handling big dataset	*	*	*	*	*	*	*			*		
Handling missing data	*	*	*	*	*	*		*	*		*	*
Handling imbalance data												
Handling noisy data	*	*	*	*	*	*			*	*		
Handling outliers								*	*		*	*
Usage on continuous data									*	*	*	
Usage on category data	*	*	*	*	*	*	*	*				*
Usage on time-series data			*			*			*	*		
Variable selection								*	*	*	*	*
Presentation												
Visualization	*	*	*	*	*	*		*			*	*
Capability												
Classification	*	*	*		*	*	*	*	*	*	*	*
Regression								*	*	*	*	*

Causal inference	*	*	*	*	*	*						
Support decision-making				*				*			*	*
Natural language processing	*	*	*		*	*	*	*	*	*		*
Image processing	*	*	*		*	*	*	*	*	*		*
Interpretation												
Explainable method	*	*	*	*	*	*	*	*			*	*
Computational requirement												
Require hardware dependency									*	*	*	
Require more training time								*	*	*		
Prone to overfitting	*	*	*	*	*	*	*	*	*	*	*	*

It is important to note that each method exhibits different strengths in different areas. In this table, the characteristics described are not meant to be exhaustive. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them.

Abbreviation: **BN** Bayesian Network, **DBN** Dynamic Bayesian Network, **DT** Decision Tree, **HMM** Hidden Markov Model, **ID** Influence Diagram, **LR** Logistic Regression, **ML** machine learning, **NB** Naïve Bayes, **NN** Neural Network, **PGM** probabilistic graphical model, **RF** Random Forest**, SVM** Support Vector Machine, **TAN** Tree Augmented Naïve Bayes, **XGBoost** Extreme Gradient Boosting.

Performance comparison of PGM and other MLs on the same dataset

Publication	Methods used	AUC	SEN	SPE	NPV	PPV
Mani et al., 2014	RF	0.57-0.65	0.82-0.94	0.18-0.47	0.28-0.73	0.55-0.70
	SVM	0.61-0.68	0.79-0.88	0.18-0.26	0.27-0.59	0.51-0.69

	KNN	0.54-0.62	0.83-0.86	0.18-0.29	0.30-0.55	0.52-0.70
	CART	0.65-0.77	0.75-0.81	0.18-0.30	0.23-0.51	0.51-0.68
	LR	0.61	0.86-0.87	0.18-0.33	0.35-0.57	0.52-0.72
	LBR	0.58-0.62	0.86-0.85	0.18-0.33	0.36-0.52	0.52-0.72
	AODE	0.53-0.61	0.85-0.88	0.18-0.36	0.38-0.54	0.52-0.73
	NB	**0.64-0.78**	**0.83-0.95**	**0.18-0.47**	**0.31-0.76**	**0.55-0.72**
	TAN	**0.53-0.59**	**0.84**	**0.18-0.32**	**0.32-0.52**	**0.50-0.72**
Gomez et al., 2019	Adaboost	0.943	0.944	0.944	0.942	0.945
	Bagged Trees	0.88	0.901	0.858	0.896	0.866
	RF	0.84	0.861	0.818	0.853	0.827
	LR	0.787	0.771	0.804	0.777	0.8
	SVM	0.755	0.641	0.868	0.705	0.831
	DT	0.751	0.816	0.687	0.788	0.726
	KNN	0.64	0.565	0.715	0.62	0.667
	NB	**0.666**	**0.431**	**0.901**	**0.61**	**0.814**
Masino et al., 2019	Adaboost	0.83-0.85	0.8	0.72	0.92	0.51
	GB	0.8-0.87	0.8	0.74	0.92	0.53
	GP	0.75-0.79	0.8	0.6	0.9	0.44
	KNN	0.73-0.79	0.8	0.55	0.9	0.39
	LR	0.83-0.85	0.8	0.74	0.93	0.52
	RF	0.82-0.86	0.8	0.74	0.92	0.53
	SVM	0.82-0.86	0.8	0.72	0.92	0.51
	NB	**0.81-0.84**	**0.8**	**0.73**	**0.92**	**0.52**
Song et al., 2020	LR	0.86	-	-	0.94-0.96	0.4-0.5
	DT	0.6-0.84	-	-	0.84-0.95	0.39-0.57
	AdaBoost	0.81-0.83	-	-	0.91-0.94	0.41-0.53

	ET	0.80	-	-	0.81-0.88	0.53-0.68
	Bagging	0.77-0.81	-	-	0.83-0.88	0.45-0.59
	RF	0.81-0.82	-	-	0.83-0.88	0.51-0.66
	GB	0.86-0.87	-	-	0.92-0.94	0.45-0.56
	GNB	**0.81-0.82**	**-**	**-**	**0.95-0.96**	**0.28-0.38**
Cabrera-Quiros et al., 2021	LR	0.79	0.78	0.8	-	0.82
	NN	0.7	0.67	0.74	-	0.73
	NB	**0.71**	**0.68**	**0.74**	**-**	**0.73**

In this table, PGM performance is highlighted in bold. Metrics presented in the table (AUC, SEN, SPE, NPV, PPV) were chosen as they were reported in the respective studies. We excluded Stanculescu et al. (2014), Honore et al. (2020), Ying et al. (2021), and Kausch et al. (2021) from this table because they only reported AUC.

Abbreviations: AODE Averaged one dependence estimators, AUC area under the curve, CART Classification and Regression Tree, DT Decision Tree, ET Extra Trees, GB Gradient Boosting, GP Gaussian Process, KNN K-Nearest Neighbour, LBR Lazy Bayesian rules, LR Logistic Regression, ML Machine Learning, NB Naïve Bayes, NN Neural Network, NPV negative predicted value, PGM Probabilistic Graphical Model, PPV positive predicted value, RF Random Forest, SEN sensitivity, SPE specificity, SVM Support Vector Machine, TAN Tree Augmented Naïve Bayes.

Changes in text: Several changes have been made in the manuscript:

In the Abstract Result section: "**When applied to the same dataset, the performance of probabilistic graphical models (area under curve: 0.53-0.84, sensitivity: 0.43-0.95, specificity: 0.18-0.90, negative predictive value: 0.31-0.96, positive predictive value: 0.28-0.81) overlapped with other machine learning models (area under curve: 0.54-0.94, sensitivity: 0.57-0.94, specificity: 0.18-0.94, negative predictive value: 0.23-0.96, positive predictive value: 0.39-0.95)**".

In the Method section, page 8, line 142-147: "**In this review, we present the performance comparison in two approaches: (1) an overall qualitative comparison between PGM and other ML methods, and (2) a deeper comparison based on studies using both PGM and other ML methods on the same dataset. Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics have been reported commonly across several studies. In addition to the performance comparison, an analysis of pediatric sepsis definitions was also conducted from selected publications."**

In the Result section, page 9, line 179-186: "**Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV:**

0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results. Furthermore, the studies that used both PGM and other ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE), whereas other attributes, such as explainability and visualization, were not examined."

The following paragraph has been added to the Discussion page 13, line 263-281: "**A comparison of the characteristics of PGM and other ML methods is presented in Table 4. Among the popular methods that are often used in the literature for PGM, we selected BN, NB, TAN, HMM, DBN, and Influence Diagram (ID). Other ML methods that we chose from our review include LR, RF, Support Vector Machine (SVM), NN, Extreme Gradient Boosting (XGBoost), and DT. We observed that all methods exhibit different strengths in different areas. For instance, several of the other ML methods shown in Table 4 are capable of performing both classification and regression, while PGM can only perform classification. The NN and the SVM excel on several criteria; however, they require longer training time, hardware dependence, and additional aids for visualization and interpretation. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them. The main disadvantage of all methods is that they are sensitive to outliers, imbalanced data, and easily prone to overfitting when the model settings are not probably figured. The complexity of the model, the training time, and the interpretability of the results are also subjected to trade-offs. It is likely that highly complex models will require a larger amount of training time and be more challenging to interpret**."

Comment 4. The discussion appears biased towards Probabilistic Graphical Models when in fact your data doesn't convincingly support or refute this assertion. For the average reader, it would be helpful to provide more information on PGM models. Also consider including: What are the advantages and disadvantages between different PGM models? How can they be implemented? What are the challenges to using them real-time? What can be done to improve their performance in the future.

Reply: Thank you for your comment. We have added an additional table (Table 4) to describe the advantages and disadvantages between different ML methods. For PGM, we included common methods described in the literature: Tree Augmented Naïve Bayes (TAN), Naïve Bayes (NB), Hidden Markov Model (HMM), Influence Diagram (ID), Bayesian Network (BN), and Dynamic Bayesian Network (DBN). For other ML methods, we included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) as these are the most commonly used methods in our review. We observe that all methods exhibit different strengths in different areas. The common drawback of all methods is that they are sensitive to outliers,

imbalance data, and are easily prone to overfitting when the model settings are not accurately figured. In addition, there are trade-offs between model complexity, training time, and interpretability. Models with a high level of complexity may require a long training period and become challenging to explain.

Changes in text: Several changes have been made in the manuscript:

We have added **Table 4** to compare the different characteristics of PGM and other ML methods.

In the Result section, page 9, line 179-186: "**Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results. Furthermore, the studies that used both PGM and other ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE), whereas other attributes, such as explainability and visualization, were not examined.**"

The following paragraph has been added to the Discussion page 13, line 263-281: "**A comparison of the characteristics of PGM and other ML methods is presented in Table 4. Among the popular methods that are often used in the literature for PGM, we selected BN, NB, TAN, HMM, DBN, and Influence Diagram (ID). Other ML methods that we chose from our review include LR, RF, Support Vector Machine (SVM), NN, Extreme Gradient Boosting (XGBoost), and DT. We observed that all methods exhibit different strengths in different areas. For instance, several of the other ML methods shown in Table 4 are capable of performing both classification and regression, while PGM can only perform classification. The NN and the SVM excel on several criteria; however, they require longer training time, hardware dependence, and additional aids for visualization and interpretation. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them. The main disadvantage of all methods is that they are sensitive to outliers, imbalanced data, and easily prone to overfitting when the model settings are not probably figured. The complexity of the model, the training time, and the interpretability of the results are also subjected to trade-offs. It is likely that highly complex models will require a larger amount of training time and be more challenging to interpret**."

Reviewer C

The article entitled 'Machine learning and probabilistic graphical model approaches in pediatric sepsis: A scoping review' aims to evaluate the feasibility of probabilistic graphical model (GPM) in pediatric sepsis diagnosis.
The topic of ML and pediatric sepsis diagnosis is of utmost importance, given its high mortality rate and the time it takes to obtain a gold-standard diagnosis through a positive blood culture. While I recognize the potential benefits of using ML models for predicting pediatric sepsis, I have significant concerns regarding the methodology and accuracy of the study in question.

Comment Line 60: 'The actual pediatric sepsis population is characterized by a complex set of characteristics. Therefore, combining different sepsis definitions might improve the prediction of sepsis in future machine learning-based studies.' Although I agree with this sentence, I don't see this as a key finding of the review.
Reply: Thank you for your comment. We have reviewed the Key findings section and removed the mentioned sentences.
Changes in text: We have removed the mentioned sentences from the Key findings.

Comment Line 60: I would add that one of the main advantages of PGMs is the capacity of handling missing data.
Reply: We agree with the reviewer that one of the main advantages of PGMs is the capability to handle missing data. We have added this point to our discussion.

Changes in text: The following paragraph has been added to the Discussion at Page 14, Line 282-284: "**In comparison with other ML methods, PGM has certain advantages and is particularly appealing with an interactive graphical representation, a wide range of methods, transparent reasoning, the ability for causal inference and handling missing data (Table 4)**". The point is also reflected in Table 4, where we compare different characteristic of PGM and other ML methods.

Table 4: Characteristic comparison between PGM methods and other ML methods.

Characteristics	**PGM Method**						**Other ML Methods**					
	TAN	NB	HMM	IDN	BNN	DBN	LR	RF	SVM	NN	XGBoost	DT
Data Handling												
Handling small data size	*	*	*	*	*	*			*			
Handling big dataset	*	*	*	*	*	*	*			*		
Handling missing data	*	*	*	*	*	*		*	*		*	*

Handling imbalance data													
Handling noisy data	*	*	*	*	*	*			*	*			
Handling outliers								*	*		*	*	
Usage on continuous data									*	*	*		
Usage on category data	*	*	*	*	*	*	*	*				*	
Usage on time-series data			*			*			*	*			
Variable selection								*	*	*	*	*	
Presentation													
Visualization	*	*	*	*	*	*		*			*	*	
Capability													
Classification	*	*	*		*	*	*	*	*	*	*	*	
Regression								*	*	*	*	*	
Causal Inference	*	*	*	*	*	*							
Support Decision-making				*				*			*	*	
Natural language processing	*	*	*		*	*	*	*	*	*		*	
Image processing	*	*	*		*	*	*	*	*	*		*	
Interpretation													
Explainable method	*	*	*	*	*	*	*	*			*	*	
Computational requirement													
Require hardware dependency									*	*	*		
Require more training time								*	*	*			

Prone to overfitting	*	*	*	*	*	*	*	*	*	*	*	*

It is important to note that each method exhibits different strengths in different areas. In this table, the characteristics described are not meant to be exhaustive. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them.

Abbreviation: **BN** Bayesian Network, **DBN** Dynamic Bayesian Network, **DT** Decision Tree, **HMM** Hidden Markov Model, **ID** Influence Diagram, **LR** Logistic Regression, **ML** machine learning, **NB** Naïve Bayes, **NN** Neural Network, **PGM** probabilistic graphical model, **RF** Random Forest**, SVM** Support Vector Machine, **TAN** Tree Augmented Naïve Bayes, **XGBoost** Extreme Gradient Boosting.

Comment Lines 91-95 –PGMs explanation could be simpler, as the reader may have difficulties to understand. PGMs are a way of presenting relationships between variables in a visual manner. These representations are constructed based on the relationships between variables. The aim is to use this representation to make predictions. Variables are represented as nodes and the relationships between them as edges. The edges are associated with probabilities, which describe the likelihood of certain events happening given certain conditions.

Reply: Thank you for your suggestion. We have rewritten the description of PGM following your suggestion to make it easier for readers to understand.

Changes in text: The paragraph has been amended at Page 12, Line 239-241: "**PGM refers to methods that use graphs (nodes, edges) and probability theory to model the association between variables, where variables are represented by nodes and their probability relationships by edges.".**

Comment Line 150: The use of AUROC as the sole evaluation metric for imbalanced datasets is also not advisable.

Reply: We agree with the reviewer that AUROC cannot be used as the sole evaluation metric for imbalanced dataset. It would be recommended to use several metrics, such as AUC, SEN, SPE, PPV, NPV, F1, G-mean, and more, at the same time to evaluate the model performance. We have rewritten the sentence to reduce misunderstanding.

Changes in text: We have amended the sentence at Page 8, Line 137-141: "**In the event of an imbalanced dataset, the use of ACC, AUC, or any single metric alone is usually not recommended because it does not accurately reflect the model's predictive ability**. **A combination of them with additional F-score, G-mean, area under precision-recall curve (AUPRC), and various other metrics that provide different views on the predicted positives and negatives should be used instead.".**

Comment Line 181: The authors suggest the use of metrics such as F1-score for evaluating ML models on imbalanced datasets. However, the study did not use those metrics to compare ML models. The authors used accuracy, sensitivity, specificity, PPV, and NPV as evaluation metrics to compare ML models. It is important to note that AUC can be affected by class imbalance and should not be used as a single metric in an unbalanced data. To compare ML models in unbalanced problems, other metrics should also be used (precision, recall, F1-score, Area under the Precision-Recall Curve).

Reply: According to the previous point, it is recommended that machine learning models be evaluated using a variety of metrics, especially in cases of imbalanced datasets. However, not all of the metrics mentioned above were reported in the studies selected. Therefore, we could only compare AUC, SEN, SPE, NPV, and PPV. We have added this point to our study's limitations.

Changes in text: The following sentences were added to the Methods section at Page 8, Line 144: "**Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics were commonly reported across several studies**". We also added the following sentences to the Study Limitation, at Page 17, Line 369-373: "**For evaluation of ML models, it is recommended to use a number of different metrics, including AUC, AUPRC, F1-score, G-mean, and more. Ideally, our performance comparison should have been conducted using these metrics. However, not all of the metrics mentioned above were reported in the selected studies. Comparing models in this manner may not represent all aspects of their performance accurately.".**

Comment: One of the main advantages of PGMs is their ability to present relationships in a visual form. To highlight the value of PGMs, the study could demonstrate how this visual representation can provide insights and improved decision-making.
Reply: Thank you for your suggestion. Figure 2 in our manuscript illustrates several examples of PGM methodologies that can be used in sepsis diagnosis. Specifically, we presented Tree Augmented Naïve Bayes (TAN), Naïve Bayes (NB), Markov Chain (MC), and Influence Diagrams (ID) in this figure. TAN and NB can be used for sepsis diagnosis or risk calculator, Markov Chain can be used for time-series data, and ID can be used for aiding clinical decision making. PGM visual representation allows clinicians to conduct inference and prediction tasks simultaneously, allowing them to diagnose as well as investigate the interactions between variables. Following your suggestion, we have added an example how this visual representation can provide insights and improved decision-making.
Changes in text: The following sentences were added to Discussion, Page 14-15, Line 298-304: ". **Figure 2 illustrates several PGM methods for sepsis applications, including TAN, NB, Markov Models, and ID for assessing sepsis, monitoring the disease's progression, and making clinical decisions. The direction and probability encapsulated in figure 2A can be examined to gain a better understanding of the causal relationship between sepsis and biomarkers as well as how each biomarker ultimately contributes to sepsis, while**

figure 2C depicts the likelihood of the patient's condition remaining in one stage or transitioning between stages".

Comment Line 222: A black-box model is one that is not transparent, making it difficult to understand how it arrived at its results. This lack of clarity does not necessarily imply that the model is unreliable, but it can raise questions about the fairness.

Reply: We agree with the reviewer that the topic of a black-box model remains controversial. At present, there is no concrete evidence to support the claim that the model is reliable or unreliable. Our perspective is that medical domains are conservative in nature and require a high degree of transparency and interpretability. It is therefore necessary to carefully consider all aspects of the black-box model until there is further evidence to support its complete reliability. We have amended our sentences to remain neutral regarding the topic.

Changes in text: The sentences have been amended in Page 11, Line 225-227: "**Despite this, we cannot deny the growing potential and achievements of deep learning methods. Therefore, it is necessary to obtain additional evidence and validation before drawing any definitive conclusions about this controversial issue.**".

Comment Line 267: Principal Component Analysis is a technique used for dimensionality reduction, but it is not required to visualize high-dimensional data. Random Forest does not have a graphical representation, however, decision trees, which are building blocks of Random Forest, can be visualized. Neural Networks (NNs) can be visualized in certain aspects using saliency maps or activation maximization. NN's is not always too complex to interpret.

Reply: Apart from the dimension reduction method, there are, indeed, other ways to visualize high-dimensional data, such as cluster analysis. In order to reflect it more accurately, we have amended our sentence. We also acknowledge that Random Forests and Neural Networks can be visualized in the way the reviewer mentioned.

Changes in text: We have amended the following paragraphs to reflect the changes at Page 14, Line 285-289: "**It may be necessary to perform a dimension reduction method or a cluster analysis in order to visualize a higher-dimensional dataset. RF and XGBoost, on the other hand, lack a graphical representation, unless they are viewed as structures of several decision trees. NN's representation is often difficult to interpret, although saliency maps and activation maximization can be used to visualize it.".**

Reviewer D

This paper provides a literature review regarding papers which used machine learning and probabilistic graphical models with the aim of predicting neonatal and pediatric sepsis.
The authors have read and reviewed the content of all papers published between the year 2000 and 2022 in the most important databases that are accessible online and that include Pubmed,

Scopus and Web of Science among others. In addition to this, the authors showed that the performances and outcomes found when using probabilistic graphical models are quite in line or only slightly inferior to the ones obtained while using machine learning or deep learning models. On the other hand, probabilistic graphical models allow for interactive graphical representations and are therefore more easily interpretable, reason why the authors suggest their use for future studies. The paper also highlights that different definitions of sepsis are used and therefore it is suggested for future authors to investigate their models based on multiple definitions.
The paper is well written and includes relevant information for future authors that might be interested in predicting sepsis in neonatal or pediatric populations. Some revisions are however suggested to improve the overall good quality of the current version of the manuscript:

1 – In the methods section the authors mentioned that they used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-Scr) checklist in order to conduct their review. Some additional information about why this checklist is relevant and was therefore chosen by the authors should be added. In addition to this the PRISMA-Scr checklist is also included in the manuscript but there is no actual reference in the Methods section to this checklist. Is this checklist supposed to be included as supplemental material only?

Reply: Thank you for your suggestion. We have added the reason why we chose PRISMA-Scr and the reference to our PRISMA-Scr checklist in the manuscript.

Changes in text: The following sentences have been added to Page 6, Line 106-109: "**This scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-Scr), which provides a systematic, transparent, and objective method for selecting, reviewing, and reporting research publications (23). The details of the PRISMA-Scr checklist for this review are included in Supplemental material.".**

Comment 2 – In the results section the authors provide information regarding the overall performance of the machine learning and probabilistic graphical models considering all the reviewed papers. It would be a nice addition for the manuscript to add just a few lines to compare the performance of the machine learning and probabilistic graphical models only considering papers that used both approaches (e.g., Mani et al., 2014, Gomez et al., 2019, Masino et al., 2019, etc.). This would allow for an even more fair comparison in terms of included patients and data.
Reply: Thank you for your recommendation. In this revised manuscript, we have presented the comparison in two approaches: (1) an overall comparison between PGM and other ML methods, and (2) a deeper comparison based on studies using both PGM and other ML methods on the same dataset.

For the first approach, we added an additional table (**Table 4**) to describe the advantages and disadvantages between different ML methods. For PGM, we included common methods

described in the literature: Tree Augmented Naïve Bayes (TAN), Naïve Bayes (NB), Hidden Markov Model (HMM), Influence Diagram (ID), Bayesian Network (BN), and Dynamic Bayesian Network (DBN). For other ML methods, we included Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) as these are the most commonly used methods in our review. We observed that all methods exhibit different strengths in different areas. The common drawback of all methods is that they are sensitive to outliers, imbalance data, and are easily prone to overfitting when the model settings are not accurately figured. In addition, there are trade-offs between model complexity, training time, and interpretability. Models with a high level of complexity may require a long training period and become challenging to explain.

In our second approach, we compared PGM and other ML methods on the same dataset. Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics have been reported across the studies. Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results.

Characteristic comparison between PGM methods and other ML methods.

Characteristics	**PGM Methods**						**Other ML Methods**					
	TAN	NB	HMM	ID	BN	DBN	LR	RF	SVM	NN	XGBoost	DT
Data Handling												
Handling small data size	*	*	*	*	*	*			*			
Handling big dataset	*	*	*	*	*	*	*			*		
Handling missing data	*	*	*	*	*	*		*	*		*	*
Handling imbalance data												
Handling noisy data	*	*	*	*	*	*			*	*		
Handling outliers								*	*		*	*
Usage on continuous data									*	*	*	

Usage on category data	*	*	*	*	*	*	*	*				*
Usage on time-series data			*			*			*	*		
Variable selection								*	*	*	*	*
Presentation												
Visualization	*	*	*	*	*	*		*			*	*
Capability												
Classification	*	*	*		*	*	*	*	*	*	*	*
Regression								*	*	*	*	*
Causal inference	*	*	*	*	*	*						
Support decision-making				*				*			*	*
Natural language processing	*	*	*		*	*	*	*	*	*		*
Image processing	*	*	*		*	*	*	*	*	*		*
Interpretation												
Explainable method	*	*	*	*	*	*	*	*			*	*
Computational requirement												
Require hardware dependency									*	*	*	
Require more training time								*	*	*		
Prone to overfitting	*	*	*	*	*	*	*	*	*	*	*	*

It is important to note that each method exhibits different strengths in different areas. In this table, the characteristics described are not meant to be exhaustive. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them.

Abbreviation: **BN** Bayesian Network, **DBN** Dynamic Bayesian Network, **DT** Decision Tree,

HMM Hidden Markov Model, **ID** Influence Diagram, **LR** Logistic Regression, **ML** machine learning, **NB** Naïve Bayes, **NN** Neural Network, **PGM** probabilistic graphical model, **RF** Random Forest**, SVM** Support Vector Machine, **TAN** Tree Augmented Naïve Bayes, **XGBoost** Extreme Gradient Boosting.

Performance comparison of PGM and other MLs on the same dataset

Publication	Methods used	AUC	SEN	SPE	NPV	PPV
Mani et al., 2014	RF	0.57-0.65	0.82-0.94	0.18-0.47	0.28-0.73	0.55-0.70
	SVM	0.61-0.68	0.79-0.88	0.18-0.26	0.27-0.59	0.51-0.69
	KNN	0.54-0.62	0.83-0.86	0.18-0.29	0.30-0.55	0.52-0.70
	CART	0.65-0.77	0.75-0.81	0.18-0.30	0.23-0.51	0.51-0.68
	LR	0.61	0.86-0.87	0.18-0.33	0.35-0.57	0.52-0.72
	LBR	0.58-0.62	0.86-0.85	0.18-0.33	0.36-0.52	0.52-0.72
	AODE	0.53-0.61	0.85-0.88	0.18-0.36	0.38-0.54	0.52-0.73
	NB	**0.64-0.78**	**0.83-0.95**	**0.18-0.47**	**0.31-0.76**	**0.55-0.72**
	TAN	**0.53-0.59**	**0.84**	**0.18-0.32**	**0.32-0.52**	**0.50-0.72**
Gomez et al., 2019	Adaboost	0.943	0.944	0.944	0.942	0.945
	Bagged Trees	0.88	0.901	0.858	0.896	0.866
	RF	0.84	0.861	0.818	0.853	0.827
	LR	0.787	0.771	0.804	0.777	0.8
	SVM	0.755	0.641	0.868	0.705	0.831
	DT	0.751	0.816	0.687	0.788	0.726
	KNN	0.64	0.565	0.715	0.62	0.667
	NB	**0.666**	**0.431**	**0.901**	**0.61**	**0.814**
Masino et al., 2019	Adaboost	0.83-0.85	0.8	0.72	0.92	0.51
	GB	0.8-0.87	0.8	0.74	0.92	0.53
	GP	0.75-0.79	0.8	0.6	0.9	0.44
	KNN	0.73-0.79	0.8	0.55	0.9	0.39

	LR	0.83-0.85	0.8	0.74	0.93	0.52
	RF	0.82-0.86	0.8	0.74	0.92	0.53
	SVM	0.82-0.86	0.8	0.72	0.92	0.51
	NB	**0.81-0.84**	**0.8**	**0.73**	**0.92**	**0.52**
Song et al., 2020	LR	0.86	-	-	0.94-0.96	0.4-0.5
	DT	0.6-0.84	-	-	0.84-0.95	0.39-0.57
	AdaBoost	0.81-0.83	-	-	0.91-0.94	0.41-0.53
	ET	0.80	-	-	0.81-0.88	0.53-0.68
	Bagging	0.77-0.81	-	-	0.83-0.88	0.45-0.59
	RF	0.81-0.82	-	-	0.83-0.88	0.51-0.66
	GB	0.86-0.87	-	-	0.92-0.94	0.45-0.56
	GNB	**0.81-0.82**	**-**	**-**	**0.95-0.96**	**0.28-0.38**
Cabrera-Quiros et al., 2021	LR	0.79	0.78	0.8	-	0.82
	NN	0.7	0.67	0.74	-	0.73
	NB	**0.71**	**0.68**	**0.74**	**-**	**0.73**

In this table, PGM performance is highlighted in bold. Metrics presented in the table (AUC, SEN, SPE, NPV, PPV) were chosen as they were reported in the respective studies. We excluded Stanculescu et al. (2014), Honore et al. (2020), Ying et al. (2021), and Kausch et al. (2021) from this table because they only reported AUC.

Abbreviations: AODE Averaged one dependence estimators, AUC area under the curve, CART Classification and Regression Tree, DT Decision Tree, ET Extra Trees, GB Gradient Boosting, GP Gaussian Process, KNN K-Nearest Neighbour, LBR Lazy Bayesian rules, LR Logistic Regression, ML Machine Learning, NB Naïve Bayes, NN Neural Network, NPV negative predicted value, PGM Probabilistic Graphical Model, PPV positive predicted value, RF Random Forest, SEN sensitivity, SPE specificity, SVM Support Vector Machine, TAN Tree Augmented Naïve Bayes.

Changes in text: Several changes have been made in the manuscript:

In the Abstract Result section: "**When applied to the same dataset, the performance of probabilistic graphical models (area under curve: 0.53-0.84, sensitivity: 0.43-0.95, specificity: 0.18-0.90, negative predictive value: 0.31-0.96, positive predictive value: 0.28-0.81) overlapped with other machine learning models (area under curve: 0.54-0.94,**

sensitivity: 0.57-0.94, specificity: 0.18-0.94, negative predictive value: 0.23-0.96, positive predictive value: 0.39-0.95)".

In the Method section, page 8, line 142-147: "**In this review, we present the performance comparison in two approaches: (1) an overall qualitative comparison between PGM and other ML methods, and (2) a deeper comparison based on studies using both PGM and other ML methods on the same dataset. Performance was assessed using AUC, SEN, SPE, NPV, and PPV, as these metrics have been reported commonly across several studies. In addition to the performance comparison, an analysis of pediatric sepsis definitions was also conducted from selected publications."**

In the Result section, page 9, line 179-186: "**Overall, the performance of PGM (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, PPV: 0.28-0.81, NPV: 0.31-0.96) vary on different settings and datasets. When comparing both approaches on the same dataset, PGM showed a comparable performance (AUC: 0.53-0.84, SEN: 0.43-0.95, SPE: 0.18-0.90, NPV: 0.31-0.96, PPV: 0.28-0.81) to other ML models (AUC: 0.53-0.94, SEN: 0.67-0.94, SPE: 0.18-0.94, NPV: 0.23-0.96, PPV: 0.39-0.95), with overlapping results. Furthermore, the studies that used both PGM and other ML methods examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE), whereas other attributes, such as explainability and visualization, were not examined."**

The following paragraph has been added to the Discussion page 13, line 263-281: "**A comparison of the characteristics of PGM and other ML methods is presented in Table 4. Among the popular methods that are often used in the literature for PGM, we selected BN, NB, TAN, HMM, DBN, and Influence Diagram (ID). Other ML methods that we chose from our review include LR, RF, Support Vector Machine (SVM), NN, Extreme Gradient Boosting (XGBoost), and DT. We observed that all methods exhibit different strengths in different areas. For instance, several of the other ML methods shown in Table 4 are capable of performing both classification and regression, while PGM can only perform classification. The NN and the SVM excel on several criteria; however, they require longer training time, hardware dependence, and additional aids for visualization and interpretation. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions. It has been demonstrated that some of the characteristics of data handling, visualization, and explainability can be overcome through the use of additional assistance, such as pre-data processing, visualization aids, and Explainable Artificial Intelligence (XAI). For instance, continuous data and time series data can be broken down into categorical and sliding-window data to use with methods that do not natively support them. The main disadvantage of all methods is that they are sensitive to outliers, imbalanced data, and easily prone to overfitting when the model settings are not probably figured. The complexity of the model, the training time, and the interpretability of the results are also subjected to trade-offs. It is likely that highly complex models will require a larger amount of training time and be more challenging to interpret**.

Comment 3 – In the discussion section they provide a full paragraph (i.e., 'Pediatric sepsis definition used in machine learning research') regarding how often the different sepsis definitions are used in the reviewed papers and grouped them together in 4 major categories. Since these definitions are however already mentioned in the results section it would be more logical to provide information about the sepsis definition and how the author group them already in the Methods section. This can be done for instance by moving part of the previously mentioned paragraph to the Methods section. In addition to this, since the authors mentioned 4 major categories used to group sepsis definitions, these 4 categories should also be indicated in Table 3.

Reply: Thank you for your suggestion. We agree that the information organized in the manner that you suggested will be more logical and easier for the readers to follow. We have moved the sepsis grouping to the Methods section and amended the sepsis definition table accordingly.

Changes in text: The following changes have been made in the manuscript:

In the Method section, page 8, line 146-147: "**In addition to the performance comparison, an analysis of pediatric sepsis definitions was also conducted from selected publications.**"

In the Result section, page 9, line 163-170: "**We analyzed the use of different pediatric sepsis definitions and classified them into four major categories: (1) positive cultures, (2) systemic inflammatory response syndrome (SIRS) with suspected or proven infections (IPCSS, 2005), (3) dysregulated infection response with organ dysfunction criteria (Adapted Sepsis-3, 2016), and (4) general infections, including bacteremia, bacterial, viral, or fungal infections (Table 2). Most studies utilized only one definition to identify the sepsis cohort, and some did not provide a rationale as to why a particular definition was chosen (n=5, 1.7%)".**

Table 2: Pediatric sepsis definition

Definitions		**Description**
IPSCC (2005)		SIRS and presence of suspected or proven infections
Positive cultures		Positive cultures of blood, CSF, etc.
Adapted Sepsis-3 (2016)		Dysregulated host response to infection and dysfunctional organs measured by age-based pSOFA
General Infections	Bacteremia	Blood stream infections with positive cultures
	Bacterial Infection	Bacterial infection with or without positive cultures
	Viral Infection	Viral infection with or without positive cultures

Fungal Infection	Invasive fungal infection with or without positive cultures

Abbreviations: **CSF** Cerebrospinal fluid, **IPSCC** International pediatric sepsis consensus conference, **SIRS** Systemic Inflammatory Response Syndrome, **pSOFA** pediatric Sequential organ failure assessment

Comment 4 – Table 1 presents both the inclusion and exclusion criteria for the title and abstract screening. However, since the exclusion criteria can be easily inferred considering the inclusion criteria, the authors can consider removing the exclusion criteria from this table.

Reply: Thank you for your suggestion. Indeed, the inclusion criteria can easily be inferred from the exclusion criteria. Therefore, we have amended the table to have only the inclusion criteria.
Changes in text: The Exclusion criteria column has been removed from Table 1.

Inclusion criteria

1. Year: 2000-2021
2. Journal articles
3. Study contains the predefined keywords
4. Study in English
5. Pediatric study
6. Study conducted on human
7. Real data use

Re-Review Comments from Reviewer D

The recent revision that was made by the authors of the manuscript with the title 'Machine learning and probabilistic graphical model approaches in pediatric sepsis: A scoping review' has highlighted some critical points in their work. This review is organized into two parts. A first one, which include critical points that led to this decision to reject this manuscript, and a second one, highlighting minor remarks.

Critical points:
Comment 1 – In order to show that the performance of probabilistic graphical models for sepsis prediction is comparable to those obtained with other machine learning approaches, the authors

reported values for different metrics which mixed the results of 5 different studies. As an example, the AUC for probabilistic graphical methods is reported to be in the range 0.53-0.84, where 0.53 was found in the study from Mani et al. and 0.84 in the study from Masino et al. This is for instance reported in the abstract (lines 49-54) and in the result section (lines 179-186). This way of combining metrics values from different studies which were performed independently by considering different populations, features and optimizations is not fair and all the comparisons should be compared for each study independently in order to reach a fair conclusion about which prediction methods have allowed to reach the best predictive results so far.

Reply: Thank you for your comment. We have amended the way we reported the performance of PGM in abstract and result section to avoid confusion for the readers. For the analysis, we did compare the performance of PGM and other ML in each study that used both approaches and reported them in Tables 3 and 4. Table 3 shows the performance of PGM and other ML for each study and Table 4 break down the performance of each method. We understand the importance to compare the performance on the same dataset as different model may perform differently on different dataset. Below is an example of a row for Table 3.

Authors, Year (Number of patients)	**Objectives**	**ML Methods**	**PGM Methods**	**Performance Metrics**	**Other ML Results**	**PGM Results**	**PGM and ML methods Performance comparison**
Gomez et al., 2019 (N=79)	To develop a minimally invasive and cost-effective tool, based on HRV monitoring and ML algorithms, to predict sepsis risk in neonates	RF LR SVM AdaBoost Bagged Trees Classification Tree K-NN	NB	SEN, SPE, PPV, NPV, AUC	SEN (0.57-0.94) SPE (0.72-0.95) PPV (0.67-0.95) NPV (0.62-0.94) AUC (0.64-0.94)	SEN (0.43) SPE (0.9) PPV (0.8) NPV (0.6) AUC (0.67)	- PGM (NB) yielded comparable SPE, PPV, NPV with other ML methods. However, the yielded NPV, SEN and AUC were lower

within
the first
48 hours
of life

Changes in text: We made the following changes
Abstract, Page 3 Line 49-50: "**When applied on the same dataset, probabilistic graphical models show a relatively inferior performance to other machine learning models in most cases".**
Result, Page 9, Line 170: "**When comparing ML and PGM models on the same dataset, PGM showed a relatively inferior performance to other ML models in most cases (Table 3 and 4)".**

Comment 2 – As a consequence to the previous point, when metrics values from each of the 5 studies that report the use of both machine learning and probabilistic graphical models are considered independently, metrics for the latter are found to be inferior in 4 out of 5 studies (with the study from Mani et al., being the exception). A careful understanding performed by the authors led to a brief consideration included in the discussion section (lines 250-251), where the authors indicated that 'In general, the performance of probabilistic graphical models, with NB being the most used model, appears relatively lower but comparable compared to other models.' This key message is however reported differently in other sections of the manuscript where the authors report an overlap in performance for the two types of models (e.g., in the abstract at lines 49-54) and can deliver a wrong idea to the reader.

Reply: Thank you for your comment. We have amended the reporting to avoid confusion for the readers.
Changes in text: We have made the following changes, Page 12, line 227-228: "**In general, the performance of PGM, with NB being the most used model, appeared relatively less efficient compared to other models"**

Comment 3 – Figure 1 addresses the fact that 69 studies were excluded from the revision process since they included wrong results. How could the authors assess that the results were wrong? In addition to this, can the author explain why they could not access 14 additional excluded studies?

Reply: There were 14 studies that we did not have access to the full-text paper even after contacting our university librarian. The 69 studies categorized under "wrong results" included studies that did not report the results with performance metrics (SEN, SPE, NPV, PPV, etc.) or studies that did not address sepsis in the results.
Changes in text: No changes in main text.

Comment 4 – The authors have not updated their literature research study with papers published after June 2022 and up to the present day. A more up to literature research should be

performed since quite some time has passed from the first submission.

Reply: We have performed an updated search again on PubMed, Scopius, and Web of Sciences up to May 2023 and included an additional 13 papers. The additional papers are highlighted in red color in Supplemental Material 1. We also have updated the manuscript accordingly regarding the additional information from these 13 papers.
Changes in text: The additional 13 papers are highlighted in red in Supplemental Material 1.

Minor remarks:
Comment 1 – The new addition table 4 does not serve the purpose in indicating why different machine learning methods can be useful in the context of sepsis prediction and should therefore not be included in the manuscript. References to table 4 in the manuscript (i.e., discussion at lines 268 and 284) can be replaced with a citation to papers that describe the characteristics of the different machine learning methods. In addition to this, also Supplementary Material 1 does not provide an additional benefit to the manuscript and only some references to previous studies using the PRISMA-ScR Checklist are more than sufficient for the current manuscript.

Reply: Thank you for your comment. The mentioned Table 4 is now Table 5 in the main text. As the table is requested by another reviewer and we think it will be helpful for the reader to have a quick reference while reading, we will keep the table in the main text.
Changes in text: No changes in main text.

Comment 2 – The authors can provide better readability to their result section by combining the two sections in which they address the use of different sepsis definition in the different papers and their classification into 4 categories. One possible way to structure this is by stating that the reviewed papers have been classified based on the use of different use of sepsis definition into four major categories: (1) positive cultures (n=…, X% of the total amount of included papers), …. In addition to this at lines 162 the authors report that study participants for one study reached 96,156 but this number is not found in any of the later tables which report the included number of patients.

Reply: As suggested by the reviewer, we have re-structured the reporting of the result section. For the number of participants, we have corrected the number.
Changes in text: The changes have been made in result section, Line 149-153:

"**We analyzed the use of different pediatric sepsis definitions and classified them into four major categories: (1) positive cultures (n=19, 26.4%), (2) systemic inflammatory response syndrome (SIRS) with suspected or proven infections (IPCSS, 2005, n=11, 15.3%), (3) dysregulated infection response with organ dysfunction criteria (Adapted Sepsis-3, 2016, n=7, 9.7%), and (4) general infections, including bacteremia, bacterial, viral, or fungal infections, (n=3, 4.2%)".**

"Number of study participants ranged from 15 to 35,074, with majority of studies focusing on infants (n=25, 34.7%). The incidence of sepsis ranged from 1.2-81%."

Comment 3 – The last part of the discussion section (lines 263-322) is extremely long and should be shortened. While it is clear that the authors wanted to emphasize the general advantages that probabilistic graphical models can hold in a prediction study, a discussion of these general advantages should not take such a big portion of the discussion section of a manuscript which focuses on the prediction of a specific medical condition.
Reply: We have summarized the discussion regarding the advantages and disadvantages of probabilistic graphical model. We have reduced the word count from 4,441 to 4,024 in the revised main text.
Changes in text: Several changes have been made and highlighted in red in the discussion section.

Comment 4 – The authors should aim at removing their personal opinion from their review study, which is for instance still present at lines 101-102 of the introduction ('as well as provide our perspective on the optimal definition of pediatric sepsis for these studies') and at lines 252-253 of the discussion ('In our view, NB may not have been the most appropriate prediction model for complicated problem, such as pediatric sepsis').
Reply: As suggested by the reviewer, we have revised the sentences to be more objective and removed personal opinions.
Changes in text: We have made changes at:

Page 6, Line 95-97: "**The result of this review will allow us to evaluate potential research opportunities to use PGM for various applications in pediatric sepsis as well as provide our perspective on the use of pediatric sepsis definition for these studies."**

Page 12, Line 247-249: "**NB may not be the most suitable prediction model for complicated medical conditions, such as pediatric sepsis".**

Comment 5 – The brief description regarding the function of NN algorithms in the discussion section (lines 205-207) are not correctly stated. All NN models require in fact both forward and back propagation, the first to make prediction on new examples and the second to adjust the weights and minimize the errors during the training of the model.
Reply: Thank you for your comment. We have corrected our sentences regarding the neural network models.
Changes in text: We have made the correction on Page 10, line 190-191: "**The weights of the nodes are adjusted by back-propagation during the training stage to minimize the error rate.**"

Comment 6 – The authors should be consistent along their manuscript in the way that they count and percentages. For instance, lines 42-45 of the abstract indicate 'Sepsis was defined

using positive microbiology cultures in 16 studies (27.1%), followed by the 2005's international pediatric sepsis consensus definition (n=11, 18.6%), and Sepsis-3 definition (n=6, 10.2%).' They should either stick to the use of x studies (y%) or (n=x, y%).

Reply: We have revised the reporting style to be consistent to (n=x, y%) as suggested by the reviewer.
Changes in text: The changes in the text are:

The result section of Abstract, Page 1, line 43-45:
We analyzed the use of different pediatric sepsis definitions and classified them into four major categories: (1) positive cultures (n=19, 26.4%), (2) systemic inflammatory response syndrome (SIRS) with suspected or proven infections (IPCSS, 2005, n=11, 15.3%), (3) dysregulated infection response with organ dysfunction criteria (Adapted Sepsis-3, 2016, n=7, 9.7%), and (4) general infections, including bacteremia, bacterial, viral, or fungal infections, (n=3, 4.2%).

Reviewer E

I read with interest the work from Nguyen et al. In the work, the authors proposed a review of the feasibility of probabilistic graph models applied in pediatric sepsis. The authors also proposed to check how pediatric sepsis is defined in the literature regarding machine learning.

I have some points that should be taken care of prior to any future publication.

In line 98, the authors define supervised and non-supervised instances of machine learning. Then, they use the word deep-learning. Deep-learning by itself is not a new learning style (non-supervised or non-unsupervised). Any deep learning application will fall into supervised or unsupervised. Please check that for writing consistency.
Reply: Thank you for your comment. We have reviewed and made amendment to the manuscript where the reviewer commented.
Changes in text: We made the following changes in the Introduction, page 5, line 78-80: "**It offers a wide range of established methods, including supervised learning (e.g., regression, classification) and unsupervised learning (e.g., clustering)"**.

In the paragraph starting in line 153, the authors mention several ways to assess the performance of a machine learning model. However, all those metrics only relate to classification tasks (a subdivision of supervised learning). Why haven't the authors mentioned ways to assess regression models? What about the 'trickier' ways to assess the performance of unsupervised models? If the authors are to compare only the performance within classification models, please make that clear in the text.
Reply: Thank you for your comment. We have added the performance evaluation for regression models and unsupervised to complete the section. We also added the related references for the readers to refer to on the topic.

Changes in text: We made the following changes in the Introduction, page 8, line 137-146: "**The most commonly used evaluation metrics for regression tasks in supervised learning are R Squares or adjusted R Squares, mean square errors (MSE) or Root mean square errors (RMSE), and mean absolute errors (MAE). These metrics measure the fit between prediction values and ground truth values. As for unsupervised learning, performance evaluation is less straightforward as it often requires the evaluation of both the results and the unsupervised algorithms employed. Essentially, it seeks to determine whether the number of clusters discovered is optimal and reliable, as well as validate whether members within a cluster and between clusters are similar. Common metrics include the Davies-Bouldin Index, Calinski-Harabasz Index, and Silhouette Coefficient (24).**"

The added references:
24.Palacio-Niño JO, Berzal F. Evaluation Metrics for Unsupervised Learning Algorithms [Internet]. arXiv; 2019 [cited 2023 Sep 7]. Available from: http://arxiv.org/abs/1905.05667

In the results, the authors first try to describe how pediatric sepsis definitions are used in the ML literature. Out of the 72 studies that met their inclusion criteria, only 25 (34%) focused on infants. The authors also identified that the most common definitions for sepsis are positive culture, SIRS, dysregulated infection, and general infections. However, I was hopping, due to the title of the paper and the proposed objectives, the authors would only try to define sepsis in infants, as this was listed in the introduction to be a current issue. I understand and appreciate the authors using the definition for sepsis as a whole, but given the objectives of this review, I would have expected to see a section on the results focusing on the subset of 25 studies whose focus was pediatric sepsis.
Reply: Thank you for your comment. We have added a brief result for the group of 25 studies on infants. Due to the heterogeneous in this group and the importance to identify sepsis infants with sepsis quicky, majority of the studies focusing on identifying sepsis children (LOS, EOS) from healthy one, or distinguishing between sepsis and other similar infections such as SIRS.
Changes in text: We added the following sentences to the result section, page 9, line 181-185: "**Among the 25 studies focusing on infants, the common sepsis definitions used were positive cultures (n=14, 56%), bacterial sepsis (n=3, 12%), time of antibiotic administration (n=3, 12%). The study objectives in this group of studies mostly focused on identifying early and late onset sepsis or distinguishing between sepsis and other signs of infection such as SIRS.**"

Also in their results, the authors provide evidence to their goal "evaluate the feasibility of PGM in pediatric sepsis application", which has underperformed compared to other ML models.
Reply: Thanks for your comment. Since these studies are ML-based and they are binary classification problem, we used performance metric as the means to evaluate the performance of PGM compared with other methods. The performance in general and the performance of PGM and other MLs on the same dataset were reported in tables 3 and 4. We mentioned in the results that PGM showed a relatively inferior performance to other MLs in most cases. However, the studies that used both PGM and other MLs examined and compared only the quantitatively measurable aspects of the methods (e.g., AUC, SEN, SPE), whereas other attributes, such as

explainability and visualization, were not examined. Therefore, we discussed these qualities in deeper manner in the discussion.
Changes in text: No change in text.

The authors bring up in **line 213** an interesting topic (i.e., explainability). Machine learning has the capacity to classify data way beyond other traditional statistical approaches, however, the higher mathematical complexity and abstraction of this models make them not trivial to interpret. The authors have even a very well-written paragraph (line 260) discussing about it.
However, especially in the health area, there is the need for these models to be as transparent as they can as they have the potential to be used for guiding decision-making processes in the area. Thus, I inquire, from the 72 shortlisted studies, what was the proportion the authors encountered of ML models having a layer of explainability? This would add more value to the paper as model explainers such as SHAP and Lime are getting more attention to unveil the 'black box' nature of such models.
Reply: Thank you for your comment. The explainability aspect is indeed important in ML-based medical studies. Unfortunately, there is a trade-off between the complexity of the model, training time, model performance, and interpretability. For simple models such as Logistic regression, Decision Tree can be fairly explained well. However, models such as Neural network required additional assistance to be explained. As you suggested, we have added a section about SHAP and LIME in the method section and report the number of studies utilized these tools in the result section.
Changes in text:
The following sentences were added to method section, page 8, line 147-155: "**The capabilities of machine learning extend far beyond those of conventional statistical methods. They are, however, difficult to interpret because of their mathematical complexity. A common question that arises when using such a model is why a particular result is reached. One way to approach this question is by examining the features involved in the learning process and the extent to which they contribute to the final result, as seen in some Explainable artificial intelligent (XAI) tools. In particular, Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have been gaining popularity due to their model-agnostic nature and user-friendly interfaces that work on several ML models. Other approaches of XAI can be referred to (25).**"
The following sentences were added to the results section, page 10, line 200-202: "**Otherwise, only three studies (34–36) utilized XAI tools, such as SHAP or LIME, to enhance the interpretation of the ML process/models.**"

In the discussion, line 276, the authors claim that there is the premature perception PGM are unreliable tools? Why? This clashes with previous assumptions the authors have made. In the introduction, the authors mentioned (line 105) "[...] amongst the ML methods, PGM [...] is one of the most robust approaches available".
Reply: In this section, we meant that the low number of PGM studies in pediatric sepsis may have given the wrong impression that the tool is not as effective as other ML methods. We have rewritten the sentences to avoid the confusion.

Changes in text: We made the changes in page 12, line 255: "**A low number of PGM studies may lead to the premature perception that PGM is an unreliable tool; however, this may not be the case as it has been shown to be effective in some precedent studies**".

Overall, the authors have a very short **result section**, with many of the results artifacts (tables and figures) being summoned only in the discussion. I kindly ask the authors to reconsider this and better describe these elements in the result section.
Reply: Thank you for your comment. We have included more information to the results section as suggested. We have added a sub-section for the 25 studies on infants and a section reported the explainability aspect in the included studies.
Changes in text: We added the following sentences to the Result section:

"**Among the 25 studies focusing on infants, the common sepsis definitions used were positive cultures (n=14, 56%), bacterial sepsis (n=3, 12%), time of antibiotic administration (n=3, 12%). The study objectives in this group of studies mostly focused on identifying early and late onset sepsis or distinguishing between sepsis and other signs of infection such as SIRS.**"
"Only three studies (34–36) utilized XAI tools, such as SHAP or LIME, to enhance the interpretation of the ML process/models."

Overall, it is always interesting to see studies that try to come up with a translational approach of machine learning, as this area of computer science generally deals with solving problems from other areas. The authors have done a great work in communicating their review in an easy-to-read format, highlighting the attention of their selected model (PGM) to its potential application in pediatric sepsis. Getting back to the **two objectives** the authors proposed to do, I see one of them as being fulfilled (i.e., evaluate the PGM feasibility). However, my biggest concern is on the unsolved objective, in which the authors proposed to define pediatric sepsis in ML studies. With the results that were presented, the authors have evidence to warrant the definition of sepsis in general (and not its pediatric 'subtype'). However, I don't see why the authors cannot do this with the data presented.
Reply: Thank you for your comment. We have considered to generalize our concept of using multiple sepsis definition in ML studies to a bigger population. However, this concept requires further validation as pediatric and adult sepsis population are quite different. We have mentioned in our Recommendation to compare single definition-learned model with the multiple definitions-learned model in order to validate our hypothesis that combining two or more sepsis definitions will improve performance of ML methods.
Changes in text: No changes in text.

Reviewer F

The authors present a scoping review on ML and PGM approaches to modeling pediatric sepsis. The main strengths of this manuscript are the thorough literature review, and the thorough

knowledge base on ML and PGM methods that the authors bring to the work. The manuscript could be improved by clarifying the purpose of the manuscript (is it a scoping review or an analysis of the feasibility of PGM?) and more closely focusing on that specific purpose.

Title: The authors could better match the title of this manuscript with the information presented: The two main points of the manuscript (see P4 L119) are 1) feasibility of PGM and 2) impact of various sepsis definitions on the use of ML for this purpose. This is not apparent from the title.
Reply: Thank you for your comment. We have amended the title to better match the two objectives of the manuscript. The revised title is "The use of probabilistic graphical models in pediatric sepsis: a feasibility and scoping review".
Changes in text: The title is changed to "**The use of probabilistic graphical models in pediatric sepsis: a feasibility and scoping review**".

Abstract & Highlight Box: As currently written, the results and conclusions of the manuscript provide different impressions. The discussion states that PGM models provided inferior performance to other ML models, however the conclusion and Key Findings state that PGM is a potentially useful method for pediatric sepsis. The authors more thoroughly describe the benefits of PGM models (interpretability, for instance) in the body and "What is known and what is new?", but this is not apparent in the abstract.
Reply: Thank you for your comment. We have amended the Key findings and abstract section to be more consistent.
Changes in text:
The following changes were made in the highlight box, key findings: "**Considering their potential qualities in explainability and transparency, PGM can be potentially useful in pediatric sepsis studies and applications"**.
The following changes were made in the abstract, conclusion: "**Current studies suggest that the performance of probabilistic graphic models is relatively inferior to other machine learning methods. However, its explainability and transparency advantages make it a potentially viable method for several pediatric sepsis studies and applications.**"

P3 L73: The second point under "What is the implication …" suggests that a lack of granular data is a flaw in previous studies. This point is not apparent from the results presented in the body of the manuscript.
Reply: This point is our proposal for the future studies, and it was not drawn from the results section. We made this proposal because different sepsis definition requires different variables, and some of the variables may not be available in the EHR that is used for the study. Additionally, it is well-known that the EHR has a very high missing data rate, therefore, there is a need to collect more granular data.
Changes in text: No changes in text.

Introduction:
P3 L83: Re-phrase to make clear that the mortality rate of infants with sepsis is 16%, not that 16% of all term infants die from sepsis.

Reply: Thank you for your comment. We have rephrased the mentioned sentences.
Changes in text: changes were made in page 5, line 67: "**Approximately 77% of infants with sepsis require intensive care, while 16% of term infants with sepsis die from this condition**".

P4 L117: The authors should provide a reference to the "past studies" referred to in the final sentence of this paragraph.
Reply: Thank you for your comment. We have added the reference for the mentioned sentences.
Changes intext: The reference is added to line 96.

Methods:
P6 L154: The authors should designate which AUC metric they are referring to (presumably the receiver-operating characteristic curve)
Reply: Thanks for your comment. We have amended the AUC to be AUROC (area under the receiver-operating characteristic curve).
Changes in text: Changes were made in page 7, line 126: "**area under the receiver-operating curve (AUROC)**".

P6 L160: It would be more accurate to state "An AUC nearer to 1.0 (in which case the ROC passes close to (0,1)) represents a higher …"
Reply: Thanks for your comment. We have amended the text.
Changes in text: We made changes to the sentence in page 7, Line 132: "**An AUROC nearer to 1.0 represents a higher capability to distinguish between positive and negative cases.**"

P6 L168: Again, this seems to suggest that the authors offer a side-by-side comparison and assessment of PGM vs other ML methods, as well as sepsis definitions - this is different from the scoping review they suggest in the title and abstract.
Reply: Thank you for your comment. We have amended the title and abstract as suggested in above comment.
Changes in text: No changes in text.

Discussion:
P8 L220: The authors have previously characterized the PGM performance as "inferior" – in Tables 3 and 4, it is clear that performance was similar between PGM and other ML models in some (but not all) cases. If the authors wish to claim that the overall performance is "reasonably well," they should indicate what data support this.
Reply: Thank you for your comment. We have amended the sentences to reflex that PGM performance is slightly inferior to other ML models in most cases.
Changes in text: we made the following changes in page 10, line 208: "**Our findings suggests that PGM performs relatively inferior compared to other ML techniques in most cases.**"

P8 L224-241: This is a relatively in-depth explanation of the function of various ML models. These details don't add much to the discussion and could be shortened.
Reply: Thank you for your comment. We have tried our best to shorten the mentioned paragraph.

Changes in text: Several changes were made in line 212-224

.

P8-9 L243-252: This part of the discussion delves into the various ML applications in pediatric sepsis – this looks more like the general scoping review proposed in the title but doesn't connect well to the two stated aims of the manuscript (is PGM feasible for pediatric sepsis, and what is the impact of varying sepsis definitions).

Reply: Thank you for your comment. In order to evaluate the feasibility of PGM, we have compared PGM with other ML methods, therefore we have briefly summarized the studies that utilized other ML methods here. We have tried our best to shorten this paragraph.

Changes in text: Several changes were made in line 225-238.

P9 L263: The "black box" problem applies to other ML tools, such as RF or boosted trees as well. The authors could also discuss the impact of model explainability tools such as LIME or SHAP here.

Reply: Thank you for your comment. We have added some analysis on the LIME and SHAP. We see that there was low number of studies utilizing LIME or SHAP to add the explainability layer to the ML models.

Changes in text:

The following sentences were added to the method section, page 8, line 147-155: "**The capabilities of machine learning extend far beyond those of conventional statistical methods. They are, however, difficult to interpret because of their mathematical complexity. A common question that arises when using such a model is why a particular result is reached. One way to approach this question is by examining the features involved in the learning process and the extent to which they contribute to the final result, as seen in some Explainable artificial intelligent (XAI) tools. In particular, Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have been gaining popularity due to their model-agnostic nature and user-friendly interfaces that work on several ML models. Other approaches of XAI can be referred to (25)."**

The following sentences were added to the discussion section, page 12, line 247-250: "**A further layer of explainability can be achieved by using additional XAI tools, such as SHAP or LIME, to enhance interpretation of the model results. This is a realistic attempt to bridge the gap between theoretical ML frameworks and practical applications."**

P9: L270: Here the authors highlight the key question: does the intuitive explainability provided by PGM justify some decrease in model performance? This is at the crux of the data they present and could be amplified throughout the paper.

Reply: Thank you for your comment. It is true to a certain degree that the explainability and transpanrency might affect the PGM model performance as discussed in the following article "Wanner, J., Herm, LV., Heinrich, K. et al. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. Electron Markets 32, 2079–2102 (2022). https://doi.org/10.1007/s12525-022-00593-5".

Changes in text: We have added this point to page 15, line 320: "**Moreover, even though explainability and transparency give PGM more advantages from other methods in ML,**

these qualities may have contributed to some degradation of its performance as well."

P10 L276: The authors assert that usage is low – the data presented in the results do not address the penetration of these methods into real-world use one way or the other.
Reply: The usage here is counted by the number of the studies that utilized the methods. Currently, for the best of our knowledge, the usage of the ML studies for real-world application remains very low as there are some concerns about reliability, transparency, reproductivity for the ML-based studies. To avoid the confusion, we have amended the mentioned sentences.
Changes in text: the changes were made in the following sentences, page 12, line 254: "**A low number of PGM studies may lead to the premature perception that PGM is an unreliable tool; however, this may not be the case as it has been shown to be effective in some precedent studies (42,43).**"

P10 L283-288: This is a nice discussion of the limitations of Naïve Bayes.
Reply: Thank you for your comment.

P10 L294: What is the source of the data in table 5? Is this the opinion of the authors? If so, this should be stated.
Reply: The table is summarized from various textbooks on machine learning topics. In this table, the characteristics described are not meant to be exhaustive. Additionally, certain limitations highlighted in some areas can be overcome through alternative solutions.
Changes in text: No change in text.

P10 L328: This is not specific to PGM – there is extensive literature on the use of other ML methods on many data types.
Reply: Thanks for your comment. We meant the transparent reasoning can be applied on top of a variety of data type. We have rewritten the sentences to make it clearer.
Changes in text: The following sentences were amended, page 14, line 195-199: "**Additionally, PGM has an extensive body of established methods that can be applied to a wide variety of data types (e.g., text, images, tabular data, and time series data), and can be tailored to meet various requirements (e.g., prediction, inference, and decision making) (18). In this way, PGM is capable of processing clinical images, physician notes, and the creation of monitoring and decision-support tools together with the causal inference ability**".

P12 L362: There are methods other than culture for identifying infections – PCR, microscopy or clinical diagnostic criteria are also available to clinicians.
Reply: Thank you for your comment. We have amended the sentence to remove the part about positive culture being the only way to confirm infection.
Changes in text: Changes in page 15, line 309.

P13 L376: It's not clear that the use of multiple sepsis definitions would improve predictive models – presumably, each sepsis definition has its own strengths or weaknesses (more specific, more general, etc.). The assertion that multiple definitions should be used appears to be an opinion and is not demonstrated by the data presented.

Reply: We propose to use multiple sepsis definition in ML studies to improve the prediction ability of the ML models and this concept requires further validation in future studies as we mentioned in the Recommendation section: "Finally, it would be desirable to conduct studies that could compare single definition-learned model with the multiple definitions-learned model in order to validate our hypothesis that combining two or more sepsis definitions will improve performance of ML methods".
Changes in text: No change in text.

Figure 3: This is a nice graphical representation of different sepsis definitions – it is unclear to me what the arrow in the lower-right labeled "sepsis" represents. (As the authors observe, a positive culture is neither necessary nor sufficient to define sepsis under most definitions)
Reply: Thank you for your comment. We have fixed the arrow position in the graph.
Changes: We have amended Figure 3.

Table 1: Typo in #6: "humans"
Reply: Thank you for your comment. We have fixed the typo.
Changes: Typo fixed in Table 1.

Reviewer G

In line 188-189: "(4) general infections, including bacteremia, bacterial, viral, or fungal infections..." was a described category for sepsis. To date, expert-derived definitions of sepsis (ICCPSOD, Sepsis-3) involve the recognition/suspicion of infection and some characterization of the pathophysiologic response. Will you clarify if any infection would be considered sepsis? Would asymptomatic culture positive infection meet criteria for general infection, and therefore sepsis in this paper?
Reply: Thank you for your question. Patient with infection has the potentially high risk to sepsis. However, there is a percentage of them turning to sepsis. Therefore, any infection, in general, would not be considered sepsis. Similarly, asymptomatic culture positive infection would not generally be considered sepsis but there are a percentage of them turning to sepsis. Our categories refer to the population that is already confirmed the status of sepsis.
Changes in text: We have added the following sentences to page 27, Line 536: "**In this review, asymptomatic patients with culture positive are not considered to have sepsis**".