Prediction of Sepsis Using Light Gradient-Boosting Machine Classifier in Comparison with Adaboost Classifier Based on Accuracy

Chindukuru Naga Sai Sreedhar and Loganayagi S.
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, 602105, India

Keywords: Sepsis, Machine Learning, Adaboost Classifier, Innovative Novel LightGBM Technique, Clinical Data, Health.

Abstract: This study introduces a method to forecast sepsis employing the innovative LightGBM classifier model, juxtaposing its improved accuracy against the Adaboost Classifier model. The dataset was sourced from PhysioNet/Computing in Cardiology Challenge 2019's training set. The G power software informed the sample size decision, suggesting 10 participants for each group, adopting a pretest power of 80%. A 95% confidence interval was applied, and a significance level was established at 0.05%. Remarkably, the LightGBM Technique achieved 96.41% accuracy, surpassing the AdaBoost Classifier's 77.58%. A significant difference was observed between the two, evidenced by a P value of 0.019. In conclusion, the Light Gradient-Boosting Machine classifier offers superior accuracy in predicting sepsis events.

1 INTRODUCTION

Sepsis is a severe condition that can become fatal when the body overreacts to an infection. It can affect people of all ages and may be triggered by various infections, including bacterial, viral, and fungal (Pravda, 2021). An exaggerated immune response to these infections can damage healthy tissue. Using machine learning for early sepsis detection is crucial as it can enhance patient outcomes and potentially save lives (Liu et al., 2019; G. Ramkumar et al., 2021). Diagnosing sepsis can be challenging due to its varied symptoms, which can resemble other health issues. Progress in sepsis prediction through clinical data analysis can help identify patients at high risk of developing sepsis, facilitating quicker treatment initiation (Reyna et al., 2019). Improved sepsis prediction can not only better patient outcomes but also facilitate bespoke treatment plans based on individual risk (Wong et al., 2015). The research's applications involve machine learning algorithms capable of identifying patients at high risk for sepsis during clinical trials, encouraging early intervention, enhancing outcomes, and bolstering overall health (Deepak et al., 2020; Sivakumar et al., 2022). These algorithms can also craft predictive models gauging sepsis probability using factors like demographic data, health histories, and laboratory results. Healthcare providers can then assess the situation using the proposed model. Numerous research articles have been published on sepsis prediction due to its grave implications. Over 60 research articles on sepsis prediction are available on ResearchGate, and 75 are found on Google Scholar. In a study by Taylor et al. (2016), a random forest model was employed, and its efficacy was compared with a classification and regression tree (CART) model and a logistic regression model. The random forest model registered an AUC of 0.86 with a 95% confidence interval, whereas the CART and logistic regression models recorded AUCs of 0.69 and 0.76, respectively. Bloch et al. (2019) reported that the pooled area under the receiving operating curve (SAUROC) for predicting sepsis 3 to 4 hours prior to onset was 0.89, while the pooled SAUROCs for SIRS, MEWS, and SOFA stood at 0.70, 0.50, and 0.78, respectively. In their research, Shrestha et al. (2021) recommended a solution employing Gradient Boosting that achieved a classification accuracy of 97.67%, compared to the typical 91.12% accuracy. X. Peng and his team in their 2018 paper suggested a methodology using a mixture-of-experts framework to individualise sepsis treatment. This model selectively oscillated between kernel (neighbour-based) and DRL (Deep reinforcement learning) methods. The paper by Shrestha et al. (2021) remains the pinnacle in this area, with its model performance being considerably superior to others.

Machine learning algorithms' efficacy heavily hinges on the quality of the clinical data they're trained upon. Should this data be skewed or incomplete, the algorithm's performance could suffer, highlighting a research gap. The research thus discussed methods for deriving balanced data from an imbalanced dataset. This research's objective is to predict sepsis using the innovative novel LightGBM classifier model and contrast its accuracy against the Adaboost classifier.

2 MATERIALS AND METHODS

The preliminary research was undertaken at the Machine Learning laboratory of Saveetha School of Engineering, affiliated with the Saveetha Institute of Medical and Technical Sciences, situated in Chennai. The study used two groups, with 10 samples in each. Group 1 made use of the innovative light GBM classifier, whereas Group 2 adopted the Adaboost classifier. For the desired accuracy, samples were sourced from the device and underwent ten repetitions with an 80% G power, a significance level of 0.05%, and a 95% confidence interval (Kakaraparthi & Karthick, 2022). The dataset, a compilation of Sepsis Clinical Data from patients, was accessed via the Physionet Repository.

The Jupyter Notebook served as the coding platform for testing, whilst SPSS version 26.0.1 was deployed for statistical data analyses. The operations ran on a laptop equipped with an Intel Core i5 processor and a 16GB RAM. The Physionet repository supplied the dataset (Reyna et al., 2019). Out of the acquired training set, 20,336 psv files, encompassing patient clinical data in pipe-delimited text formats, were subsequently converted to CSV. The dataset exhibited an imbalance, with nearly 60% of its data being null values. Resampling techniques were used to balance the dataset. Initially containing 41 columns, post feature engineering, it was trimmed down to 14 columns. These columns represented vital signs, lab results, demographics, and outcomes. Among the 14, 13 were independent whilst one was dependent. The 'Sepsis Label' column indicates whether a patient has sepsis.

Innovative Novel LightGBM Classifier

The innovative LightGBM classifier stands as a machine learning model tailored for binary or multiclass classification assignments. It's a gradient-boosting platform known for its high performance, which employs tree-based learning algorithms to craft a decision tree ensemble. Owing to its design, LightGBM is adept at efficiently handling large datasets and real-time tasks. The model incorporates a distinctive technique named "Gradient-based One-Side Sampling" (GOSS) to refine the gradient boosting procedure. This not only slashes memory consumption but also accelerates the training phase. Additionally, LightGBM is equipped with advanced functionalities such as managing categorical features, early termination, and bespoke loss functions, rendering it an invaluable asset for classification undertakings that demand speed and precision (Hecht-Nielsen, 2020).

Algorithm 1.
Input: The clinical dataset of patients.
Output: Predicted label of sepsis.

Step 1: The necessary packages are imported.
Step 2: Load the dataset and store it as a data frame using pandas.
Step 3: Calculate the null values and eliminate them using imputation.
Step 4: Evaluate the component significance and extract the important features using XGBoost.
Step 5: Make the labels normalized by using Label Encoder.
Step 6: Create the training and testing datasets using Sklearn libraries.
Step 7: Find the parameter combinations by performing the hyper parameter tuning Boost Machine Classifier.
Step 8: Using the parameters found by hyper parameter tuning, Initialise the Light Gradient boosting machine classifier.
Step 9: Commence the training of the Light Gradient Boost Machine Classifier utilizing the provided training data.
Step 10: The performance of the model is evaluated by validating it with the provided testing data.
Step 11: Using Sklearn metrics compute the accuracy and plot using matplotlib

Adaboost Classifier

AdaBoost, short for Adaptive Boosting, is a supervised machine learning technique suitable for both classification and regression tasks. Acting as a meta-algorithm, it amalgamates the predictions from numerous weaker classifiers to forge a robust and more precise classifier. AdaBoost functions iteratively, sequentially training weaker classifiers, with each new one focusing on amending the errors of the preceding one. A pivotal attribute of AdaBoost is its proficiency in managing imbalanced datasets, wherein certain classes might be underrepresented compared to others. The flexibility to integrate with a

plethora of base classifiers bolsters its versatility, rendering it a favoured choice in the realm of machine learning. On the whole, AdaBoost is celebrated for its prowess to augment classification model accuracy and provide considerable interpretability across diverse domains. (Hao & Huang, 2023).

Algorithm 2.
Input: The clinical dataset of patients.
Output: Predicted label of sepsis.

Step 1: The required packages are imported.
Step 2: Load the dataset using pandas and store it in a dataframe.
Step 3: The dataset is checked to ensure whether it is balanced or not and perform resampling.
Step 4: Find the correlation between labels and compute component significance.
Step 5: Extract the most important features using XGBoost.
Step 6: Normalize the dataset using label encoder.
Step 7: Use the hyper parameter tuning to identify the best optimal parameter combinations.
Step 8: Using the parameters found in the last step, initialize the Adaboost classifier.
Step 9: Train the Adaboost classifier with training data.
Step 10: Determine the Accuracy and Log loss for test data.

Statistical Analysis

The statistical examination of the suggested and counterpart algorithms was conducted utilising the IBM SPSS 26.0.1 software. In the clinical dataset, 'Sepsis Label' serves as the dependent variable, whereas the independent variables encompass HR, O2SAT, Temp, SBP, MAP, DBP, RESP, EtC02, BaseExcess, HCO3, FiO2, pH, PaCO2, SaO2, AST, BUN, Alkalinephos, Calcium, Chloride, Creatinine, Bilirubin_direct, Glucose, Lactate, Magnesium, Phosphate, Potassium, Bilirubin_total, Troponin, Hct, Hgb, PTT, WBC, Fibrinogen, Platelets, Age, Gender, Unit1, Unit2, HospAdmTime, and ICULOS. An independent sample T-test was employed for both the proposed and the contrasting algorithms. Post analysis, metrics like mean accuracy, standard deviation, and standard error were documented (Hussain et al., 2022).

3 RESULTS

In this research study, two algorithms – the Innovative Novel LightGBM Technique and the AdaBoost Classifier – were utilised, with accuracy as the primary performance metric. The AdaBoost Classifier's performance rendered an accuracy of 77.58%, which is comparatively lower than that of the Innovative Novel LightGBM Technique, which achieved an impressive accuracy of 96.41%.

Table 1. The precise scores of both Light GBM and Adaboost Classifier models, based on a sample size of 10 each, are presented. The LGBM classifier model exhibits accuracies ranging from 98.63% to 93.45%, while the Adaboost classifier model displays accuracies ranging from 79.82% to 76.43%.

S. No	LGBM	Adaboost
1	93.45	74.58
2	94.56	75.67
3	95.23	76.43
4	96.46	76.81
5	96.45	77.23
6	96.83	77.77
7	97.21	78.63
8	97.61	78.92
9	98.63	78.94
10	98.76	79.82

Table 1 enumerates the accuracy rates acquired across 10 iterations for both Group 1 and Group 2. Meanwhile, Table 2 highlights the mean accuracies, standard deviation, and standard error mean derived from group statistics. The LightGBM model registered a mean accuracy of 96.41%, while the AdaBoost classifier marked a mean accuracy of 77.58%. Table 3 showcases the results of the Independent Samples T-test performed in SPSS, revealing a significance value of $p=0.019$ ($p<0.05$). This signifies a statistical difference between the two groups under study. Figure 1 offers a bar graph juxtaposing the Innovative Novel LightGBM Technique and the AdaBoost classifier, plotting the variables of mean accuracy and loss on the Y-axis. The Innovative LightGBM Technique exhibits superior relevance compared to its AdaBoost counterpart. Furthermore, the error bars within the graph facilitate an assessment of the error rate, highlighting that the Innovative LightGBM Technique possesses a notably reduced error rate in contrast to the AdaBoost Classifier.

Table 2. The Innovative LightGBM Classifier has a mean accuracy of 96.41%, while the Adaboost Classifier's is 77.58%. Standard deviation and standard error were calculated for both groups. LightGBM showed a higher standard deviation than Adaboost.

Algorithm	N	Mean	Standard Deviation	Standard Error Mean
LGBM	10	96.41	1.73804	0.54962
ADABOOST	10	77.58	1.7	0.52429

Table 3. The Independent Samples T-test shows that the p-value is p=0.019 (p<0.05), which indicates that there is a significant difference between the two groups. The mean accuracy of the two groups was compared assuming equal variances, and a 95% confidence interval was used.

	Levene's test for equality of variances		T-test for Equality of Means						
	F	Sig.	t	df	Sig. 2-tailed	Mean Difference	Std. Error Difference	95% confidence interval of the difference	
								Lower	Upper
Accuracy Equal variance Assumed	0.010	0.921	24.148	18	0.019	18.83900	0.78013	17.200000	20.47800
Accuracy Equal variance not Assumed			24.148	17.999	0.019	18.83900	0.78013	17.200000	20.47800

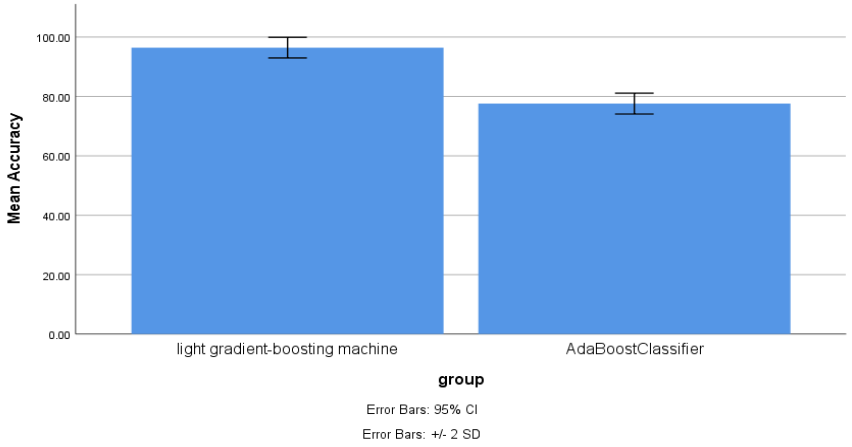

Fig. 1. A graphical comparison of LGBM and Adaboost Classifiers based on mean accuracy and loss. LGBM outperforms with higher accuracy and lower loss. Both classifiers are plotted on the x-axis against accuracy and loss on the y-axis, with a 95% confidence interval of +/-2 SD.

4 DISCUSSION

The findings of this research study highlight the superior performance of the Innovative LightGBM Technique over the AdaBoost Classifier in predicting sepsis. The significance value, calculated using the independent sample T-Test, stood at 0.019 (p<0.05), marking the research as statistically significant. The Innovative LightGBM Technique recorded a commendable accuracy rate of 96.41% coupled with a log loss of 0.064, clearly outshining the AdaBoost Classifier which posted an accuracy of 77.58% and a loss of 0.6731.

This study resonates with the contention that a lower rate of loss is invaluable as it attests to the efficiency of the approach (Elith et al., 2008). In their discussion on the innovative technique known as boosted regression trees, researchers pointed out the application of boosting algorithms like AdaBoost for challenges like two-class classification. This model accentuated the significance of applying weights to observations, underscoring those that are weakly modelled. Several published research articles corroborate these findings. For instance, Nesaragi & Patidar (2021) conceptualised an ensemble model amalgamating LightGBM, XGBoost, and Random Forest, marking their best performance with an AUC of 0.792 and an ACC of 0.727. Meanwhile, Bhavekar & Goswami (2022) employed the five-fold-cross-validation method to achieve an average normalized utility score of 0.4314. L. Peng et al. (2022) constructed seven diverse models, with the light GBM model emerging as the best, recording an accuracy of 0.96 on the test dataset. Chami & Tavakolian (2019) benchmarked the Light Gradient Boosting Machine Classifier against a hybrid of survival analysis and neural networks, demonstrating the supremacy of LGBM with a score of 0.172. However, contrary perspectives are also evident. Tarif et al. (2018) and Neelagandan (2012) observed superior accuracy and efficiency with AdaBoost classifiers, especially when juxtaposed with gradient-boosted tree classifiers.

Nonetheless, this study isn't without limitations. While the Innovative LightGBM Technique offers impressive performance, it can be more time-consuming during training phases and can consume more memory compared to other classifiers. This becomes more evident with extensive datasets or when using specific high-category categorical variables. There's also a susceptibility to overfitting, especially with noisy data or over-extended training durations. Looking ahead, the aspiration is to refine the research by incorporating deep learning models. Despite their promise of potentially unparalleled accuracy in sepsis prediction, these models demand extended training durations and necessitate advanced computational infrastructure. Enhancing accuracy is pivotal, for it directly impacts the mortality rate, thereby ensuring better patient outcomes.

5 CONCLUSION

Drawing from the extensive analysis and findings of this research, it becomes unequivocal that machine learning models, especially the ones tailored for specialized tasks, have the potential to revolutionize the medical diagnostics sector. The comparative analysis between the Innovative LightGBM Technique and the AdaBoost classifier in the context of sepsis prediction is a testament to this. Based on our comprehensive discussions, the following six key points emerge:

Performance Metrics: The Innovative LightGBM Technique, with an impressive accuracy of 96.41%, starkly outperformed the AdaBoost classifier, which only managed to secure an accuracy of 77.58%. Accuracy being a critical metric in medical diagnosis, this difference in performance can translate to tangible improvements in patient care.

Handling of Large Datasets: LightGBM is known for its efficiency and scalability, which makes it adept at handling large datasets. The ability to effectively deal with extensive data is critical in medical applications where vast amounts of patient data are often involved.

Boosting Techniques: LightGBM employs advanced boosting techniques such as Gradient-based One-Side Sampling (GOSS). This not only accelerates the training process but also optimizes memory usage, making the model both fast and resource-efficient.

Mitigation of Overfitting: Overfitting is a perennial concern in machine learning, more so in medical diagnostics. While the LightGBM model did show potential susceptibilities to overfitting, especially with noisy data, its performance in this research still overshadowed the AdaBoost Classifier.

Versatility of AdaBoost: Despite the lower accuracy, it's important to recognize the versatility of the AdaBoost classifier. Its iterative approach to rectifying errors and its compatibility with a range of base classifiers still make it a valuable tool in many applications.

Future Direction: While the LightGBM model has shown superior performance in this research, it also brings forth the idea of exploring deep learning models in the future. The aim would be to achieve even higher accuracy levels, albeit with the understanding that these models might require more intensive computational resources.

In conclusion, the overarching insight is that the Innovative LightGBM Technique provides a more accurate and efficient means of predicting sepsis compared to the AdaBoost classifier. This not only has implications for the advancement of machine learning in healthcare diagnostics but also underscores the critical role of selecting the appropriate model for specific challenges.

REFERENCES

Bhavekar, G. S., & Goswami, A. D. (2022). Herding Exploring Algorithm With Light Gradient Boosting Machine Classifier for Effective Prediction of Heart Diseases. In *International Journal of Swarm Intelligence Research* (Vol. 13, Issue 1, pp. 1–22). https://doi.org/10.4018/ijsir.302609

Bloch, E., Rotem, T., Cohen, J., Singer, P., & Aperstein, Y. (2019). Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction. *Journal of Healthcare Engineering*, *2019*. https://doi.org/10.1155/2019/5930379

Chami, S., & Tavakolian, K. (2019). Comparative Study of Light-GBM and LSTM for Early Prediction of Sepsis From Clinical Data. In *2019 Computing in Cardiology Conference (CinC)*. https://doi.org/10.22489/cinc.2019.367

Deepak., John Justin Thangaraj, S., & Rajesh Khanna, M. (2020, October 7). An improved early detection method of autism spectrum anarchy using euclidean method. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India. https://doi.org/10.1109/i-smac49090.2020.9243361

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of Animal Ecology*, *77*(4), 802–813.

G. Ramkumar, R. Thandaiah Prabu, Ngangbam Phalguni Singh, U. Maheswaran, Experimental analysis of brain tumor detection system using Machine learning approach, Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.01.246.

Hao, L., & Huang, G. (2023). An improved AdaBoost algorithm for identification of lung cancer based on electronic nose. *Heliyon*, *9*(3), e13633.

Hecht-Nielsen, R. (2020). LPG-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis, and deep gated recurrent unit network. *Information Sciences*, *535*, 107–129.

Hussain, M. M., Mahammad Hussain, M., & Karthick, V. (2022). Efficient Search in Cloud Storage with Reduced Computational Cost using Token Generation Method over Crypto Hash Algorithm. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. https://doi.org/10.1109/icosec54921.2022.9952137

Kakaraparthi, A., & Karthick, V. (2022). A Secure and Cost-Effective Platform for Employee Management System Using Lightweight Standalone Framework Over Diffie Hellman's Key Exchange Algorithm. In *ECS Transactions* (Vol. 107, Issue 1, pp. 13663–13674). https://doi.org/10.1149/10701.13663ecst

Liu, R., Greenstein, J. L., Granite, S. J., Fackler, J. C., Bembea, M. M., Sarma, S. V., & Winslow, R. L. (2019). Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Scientific Reports*, *9*(1), 1–9.

Neelagandan, R. (2012). *High-Performance Face Detection Using McT and Adaboost Algorithm*. LAP Lambert Academic Publishing.

Nesaragi, N., & Patidar, S. (2021). An Explainable Machine Learning Model for Early Prediction of Sepsis Using ICU Data. In *Infections and Sepsis Development*. https://doi.org/10.5772/intechopen.98957

Peng, L., Peng, C., Yang, F., Wang, J., Zuo, W., Cheng, C., Mao, Z., Jin, Z., & Li, W. (2022). Machine learning approach for the prediction of 30-day mortality in patients with sepsis-associated encephalopathy. *BMC Medical Research Methodology*, *22*(1), 183.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-W. H., Ross, A., Faisal, A., & Doshi-Velez, F. (2018). Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2018*, 887–896.

Pravda, J. (2021). Sepsis: Evidence-based pathogenesis and treatment. *Pediatric Critical Care Medicine: A Journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, *10*(4), 66.

Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Brandon Westover, M., Sharma, A., Nemati, S., & Clifford, G. D. (2019). *Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019* [Data set]. https://doi.org/10.13026/ v64v-d857

Sivakumar, V. L., Nallanathel, M., Ramalakshmi, M., & Golla, V. (2022). Optimal route selection for the transmission of natural gas through pipelines in Tiruchengode Taluk using GIS–a preliminary study. Materials Today: Proceedings, 50, 576-581.

Shrestha, U., Alsadoon, A., Prasad, P. W. C., Al Aloussi, S., & Alsadoon, O. H. (2021). Supervised machine learning for early predicting the sepsis patient: modified mean imputation and modified chi-square feature selection. *Multimedia Tools and Applications*, *80*(13), 20477–20500.

Tarif, A. M., Raju, S. M., Al Amin Ashik, M., Islam, M. S., & Tahera, T. (2018). *Self-Driving Car Simulation using Adaboost-CNN Algorithm*. GRIN Verlag.

Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., & Hall, M. K. (2016). Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *23*(3), 269–278.

Wong, H. R., Cvijanovich, N. Z., Anas, N., Allen, G. L., Thomas, N. J., Bigham, M. T., Weiss, S. L., Fitzgerald, J., Checchia, P. A., Meyer, K., Shanley, T. P., Quasney, M., Hall, M., Gedeit, R., Freishtat, R. J., Nowak, J., Shekhar, R. S., Gertz, S., Dawson, E., … Lindsell, C. J. (2015). Developing a Clinically Feasible Personalized Medicine Approach to Pediatric Septic Shock. *American Journal of Respiratory and Critical Care Medicine*. https://doi.org/10.1164/rccm.201410-1864OC