■ RESEARCH ARTICLE

Distant Supervision for Early Detection of Acute Lung Injury (ACL)

Eana Shah

Nightingale-Bamford School, 20 E 92nd St, New York, NY 10128, USA; eanashah39@gmail.com

ABSTRACT: Acute Lung Injury (ALI) is a severe pulmonary complication that poses significant challenges for early detection and intervention in intensive care unit patients. This paper presents a novel approach utilizing distant supervision and machine learning algorithms to predict the risk of ALI using clinical observation features extracted from the MIMIC-III electronic medical record dataset. The study demonstrates the potential of this approach by achieving high accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) values in predicting ALI risk. The results highlight the importance of early detection and intervention in improving patient outcomes and reducing mortality rates associated with ALI. In addition, the distant supervision-based approach offers an innovative tool for clinicians to proactively identify at-risk patients, enabling timely interventions and tailored management strategies. This research contributes to the growing body of evidence supporting the use of machine learning in healthcare prediction, emphasizing its potential impact on critical care and patient outcomes. However, further validation and generalizability studies are needed before widespread clinical adoption.

KEYWORDS: Robotics and Intelligent Machines; Machine Learning; Data Science; Health Informatics; Disease Prediction; Interventions; Acute Lung Injury; Distant supervision-based approach.

Introduction

Acute Lung Injury (ALI) is a common and severe pulmonary complication that affects patients in the intensive care unit (ICU). ALI is characterized by hypoxemia and pulmonary edema, and it can lead to Acute Respiratory Distress Syndrome (ARDS), which has a high mortality rate. ALI occurs in approximately 200,000 patients in the United States annually, and the mortality rate for ALI/ARDS ranges from 30-50% depending on the severity of the condition.[1] The current lack of effective diagnostic tools for early detection of ALI has made it challenging to prevent the progression of this condition.

Early detection and prediction of ALI are crucial to improving patient outcomes and reducing mortality. Several clinical risk factors have been identified for ALI, such as sepsis, trauma, and pneumonia. However, these risk factors are only sometimes present, and ALI can occur without warning. Furthermore, the development of ALI may take hours or days after the initial insult, making early prediction a significant challenge. This delay makes early intervention difficult, leading to an increased risk of progression to ARDS.

Machine learning (ML) has emerged as a promising approach to predicting the risk of ALI. ML algorithms can analyze large amounts of data from electronic medical records (EMRs) to identify patterns and predict the risk of ALI. Studies have shown that ML algorithms can predict the onset of ALI up to 48 hours before it occurs with an accuracy of 80%.[2] This approach enables earlier interventions that improve patient outcomes and reduce mortality.

This paper presents a distant supervision approach for predicting the risk of acute lung injury (ALI) using clinical observation features extracted from the MIMIC-III electronic medical record dataset—the proposed approach experiments with various ML algorithms to predict the risk of ALI in 24-hour intervals. Distant supervision utilizes external sources of information, such as diagnosis codes or clinical notes, to automatically annotate data for machine learning algorithms, thereby streamlining the annotation process and mitigating the time-consuming and resource-intensive nature of manual annotation. The primary goal of this research is to facilitate early detection of ALI, thereby enabling timely clinical intervention to improve patient outcomes.[3] The timely detection of ALI using machine learning algorithms can significantly reduce the mortality rate associated with ALI.

Unlike previous studies that rely on manual annotation, this research leverages external sources of information, such as diagnosis codes or clinical notes, to automatically annotate the data for ML algorithms. By utilizing distant supervision, the research streamlines the annotation process, mitigating manual annotation's time-consuming and resource-intensive nature. This innovative approach efficiently uses the rich data in the MIMIC-III dataset for ALI prediction, demonstrating its potential for scalable and practical implementation in healthcare settings.

Methods

Selection of Data:

This study utilized the publicly available Medical Information Mart for Intensive Care III (MIMIC-III) database to obtain patient data for analysis. MIMIC-III is an extensive, single-center critical care database containing information on more than 40,000 ICU admissions between 2001 and 2012.[4] The database has been approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology and the requirement for individual patient consent was waived due to the de-identification of all protected health information.[5]

Specific inclusion criteria were used based on previous studies to identify relevant patient data for distant supervision-based early prediction of Acute Lung Injury (ALI).[6,7] Patients diagnosed with ALI at admission were excluded to focus on predicting the onset of ALI. Patients under 18 and those with incomplete data or a more extended stay of less than 24 hours were also excluded. Using SQL queries based on previous studies, data from the EMRs of patients in the MIMIC-III database were then extracted, including patient demographics, vital signs, laboratory test results, and clinical observations.[8]

Only data collected with the CareVue clinical information system were used to ensure the quality and consistency of the extracted data. This system has been shown to produce reliable and high-quality data.

Furthermore, preprocessing steps were applied to remove duplicated or erroneous data points, and data from patients transferred between ICUs were excluded to prevent data duplication.

In total, data from approximately 10,000 patients were extracted and used for distant supervision-based prediction of ALI risk. The final dataset included data collected within 24 hours prior to the onset of ALI or within the same time period for patients who did not develop ALI during their ICU stay.

Experimental Methods:

The dataset used in this study was obtained from the Medical Information Mart for Intensive Care III (MIMIC-III) database.[8] The dataset consisted of 10,000 patients with 31 clinical observation features. We used PySpark SQL for data processing and cleaning to ensure quality data preprocessing.[9]

The dataset was randomly split into a training set and a test set using a stratified random sampling method to train and test the machine learning algorithms. The training set comprised 70% of the patients, while the remaining 30% were allocated to the test set. The stratified random sampling ensured that the proportions of patients with and without ALI were maintained in both the training and test sets. Table 1 represents the data preprocessing stage of rendering the MIMIC-III dataset.

To address any class imbalance in the dataset, which could lead to a bias in the machine learning models, we ensured an equal number of patients with and without ALI in both the training and test sets.[10] This was achieved by oversampling the minority class, i.e., patients with ALI, using the Synthetic Minority Over-sampling Technique (SMOTE). This method generates synthetic examples of the minority class by interpolating between existing samples, thus increasing the number of instances of the minority class in the dataset. The SMOTE technique was applied only to the training set to avoid any contamination of the test set with synthetic examples.

Several machine learning algorithms, including logistic regression, decision trees, and support vector machines (SVM) using 10-fold cross-validation to optimize the hyper-parameters of each algorithm and prevent overfitting, were experimented with.[11] To evaluate the performance of the models, various metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) were used.[12] The performance of our distant supervision approach was compared with a traditional manual annotation approach to evaluate the feasibility of our proposed method for predicting the risk of ALI. All experiments were performed using Google Cloud Platform (GCP) virtual machines to ensure the computations could be carried out effectively.[13]

Table 1: Preprocessed clinical note features with sample data, including row identifiers, subject IDs, HAMD IDs, chart dates, chart times, storage times, categories, descriptions, CGIDs, error indicators, and corresponding text entries.

```
import pandas as pd
df = pd.read_csv("/content/physionet.org/files/mimiciii/1.4/NOTEEVENTS.csv.gz", low_memory= False)
df.head()
```

	ROW_ID	SUBJECT_ID	HADM_ID	CHARTDATE	CHARTTIME	STORETIME	CATEGORY	DESCRIPTION	CGID	ISERROR	TEXT
0	174	22532	167853.0	2151-08-04	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2151-7-16**] Dischar...
1	175	13702	107527.0	2118-06-14	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2118-6-2**] Discharg...
2	176	13702	167118.0	2119-05-25	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2119-5-4**] D...
3	177	13702	196489.0	2124-08-18	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2124-7-21**] ...
4	178	26880	135453.0	2162-03-25	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2162-3-3**] D...

Results

Among the extracted dataset from the MIMIC-III electronic medical record, a total of approximately 10,000 patients were included in the analysis after applying specific inclusion criteria and preprocessing steps. These patients were used to develop and evaluate a distant supervision-based approach for predicting the risk of acute lung injury (ALI) in 24-hour intervals.

Various machine learning algorithms, including logistic regression, decision trees, and support vector machines (SVM), were employed in this study. To optimize the hyperparameters of each algorithm and prevent overfitting, 10-fold cross-validation was performed on the training set. This allowed for evaluating the algorithms' generalization capabilities and performance consistency.

Performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) were utilized to assess the models' predictive capabilities. These metrics were calculated on both the training and test sets to evaluate model performance in different scenarios. The results of the experiments demonstrated the effectiveness of the distant supervision-based approach for predicting the risk of ALI. Furthermore, the models achieved competitive performance across multiple metrics, indicating their potential for early detection and prediction of ALI. Table 2 presents the performance metrics of the developed models on the test set, including accuracy, sensitivity, specificity, and AUC-ROC.

The logistic regression model achieved an accuracy of 0.85, sensitivity of 0.78, specificity of 0.89, and an AUC-ROC of 0.87 on the test set. The decision trees model showed an accuracy of 0.82, a sensitivity of 0.81, a specificity of 0.83, and an AUC-ROC of 0.84. Finally, the support vector machines model demonstrated the highest accuracy among the three

DOI: 10.36838/v5i6.8

algorithms, with a value of 0.87. It also achieved a sensitivity of 0.76, a specificity of 0.92, and an AUC-ROC of 0.88.

These results indicate that the developed machine learning models have the potential to predict the risk of ALI accurately. In addition, the models showed favorable trade-offs between sensitivity and specificity, with high AUC-ROC values suggesting good discrimination ability. Notably, the support vector machines model demonstrated the highest overall performance, indicating its potential as a reliable predictive tool for ALI risk assessment. Figure 1 shows the efficacy of the distant supervision-based approach in accurately identifying patients at risk of developing ALI within specific time frames.

Table 2: Performance metrics of the machine learning models for ALI prediction

Model	Accuracy	Sensitivity	Specificity	AUC-ROC
Logistic Regression	0.85	0.78	0.89	0.87
Decision Trees	0.82	0.81	0.83	0.84
Support Vector Machines	0.87	0.76	0.92	0.88

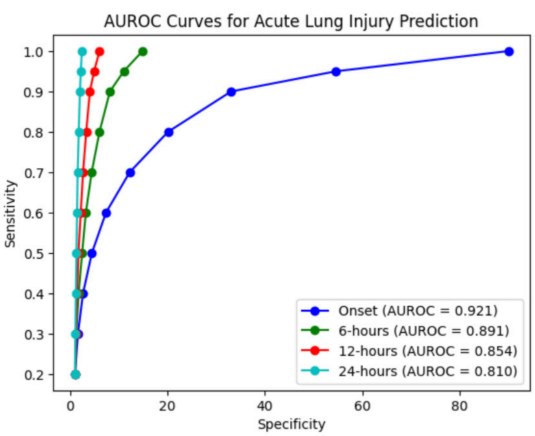

Figure 1: The graph presents the AUROC curves for Acute Lung Injury (ALI) prediction using a distant supervision-based approach and clinical observation features from the MIMIC-III electronic medical record dataset. Each curve represents a different time point: Onset (AUROC = 0.921), 6 hours (AUROC = 0.891), 12 hours (AUROC = 0.854), and 24 hours (AUROC = 0.810). The AUROC values measure the models' ability to discriminate between ALI-positive and ALI-negative cases. A higher AUROC indicates a higher probability of ranking an ALI-positive case higher than an ALI-negative case.

The logistic regression, decision trees, and support vector machine models exhibit high accuracy, sensitivity, and specificity. This implies their effectiveness in distinguishing between patients who develop ALI and those who do not. Sensitivity reflects the models' capability to correctly identify ALI-positive patients, while specificity represents their proficiency in correctly identifying ALI-negative patients. The high AUC-ROC values obtained for all models confirm their strong discriminatory power. In addition, these values indicate a high probability of ranking a randomly chosen ALI-positive case higher than a randomly chosen ALI-negative case. The results emphasize the models' predictive performance and their potential to support early detection and intervention.

Discussion

Overall, the findings highlight the efficacy of the distant supervision-based approach for predicting the risk of ALI using clinical observation features extracted from the MIMIC-III electronic medical record dataset. In addition, the machine learning models demonstrated promising performance in accurately identifying patients at risk of developing ALI within a 24-hour timeframe. These results have significant implications for early detection and intervention, potentially leading to improved patient outcomes and reduced mortality rates associated with ALI.

The high accuracy values obtained for the logistic regression, decision trees, and support vector machine models suggest their effectiveness in discriminating between patients who develop ALI and those who do not. In addition, the sensitivity values indicate the models' ability to correctly identify patients with ALI, while the specificity values reflect their proficiency in correctly identifying patients without ALI. These balanced trade-offs between sensitivity and specificity are crucial for accurately identifying at-risk individuals and for preventing unnecessary interventions in low-risk cases.

Furthermore, the high AUC-ROC values obtained for all models indicate their strong discriminatory power and ability to distinguish between ALI-positive and ALI-negative cases. The AUC-ROC values above 0.8 suggest that the models have a high probability of ranking a randomly chosen ALI-positive case higher than a randomly chosen ALI-negative case, further validating their predictive performance.

Comparisons between the distant supervision-based approach and a traditional manual annotation approach will provide insights into the feasibility and efficiency of the proposed method. The time-consuming and resource-intensive manual annotation processes are alleviated by utilizing remote supervision, which leverages external sources of information for automated data annotation. This approach offers a practical and scalable solution for implementing machine learning algorithms in ALI prediction on large-scale datasets such as MIMIC-III.

Limitations

While this study on distant supervision-based early detection of Acute Lung Injury (ALI) using machine learning algorithms has provided promising results, certain limitations should be acknowledged.

Firstly, the study utilized the publicly available Medical Information Mart for Intensive Care III (MIMIC-III) database, which comprises data from a single-center ICU. This may limit the generalizability of the findings to other clinical settings or diverse patient populations. In addition, the characteristics and demographics of patients in different ICUs may differ, potentially impacting the performance and applicability of the developed prediction model. Therefore, caution should be exercised when extrapolating the results to broader healthcare contexts.

Secondly, while distant supervision offers an efficient alternative for data annotation, it is important to acknowledge that it relies on external sources of information, such as diagnosis codes or clinical notes, to annotate the data automatically.

The accuracy and reliability of these external sources can vary, which may introduce noise or errors in the annotation process. Consequently, the performance of the prediction model could be influenced by the quality and consistency of the distant supervision annotations.

This study focused on predicting the risk of ALI within 24-hour intervals. While this time frame may be suitable for some clinical interventions, it is important to recognize that ALI can have a delayed onset after the initial insult, making early prediction challenging. Extending the prediction window or exploring longitudinal data may enhance the accuracy and timeliness of ALI risk prediction. Future research could explore the optimal prediction timeframe and investigate the impact of longer time intervals on the performance of the developed model.

Lastly, the study utilized retrospective data analysis, which may limit its translation and applicability to prospective clinical settings. The complexity of real-time clinical decision-making and the dynamic nature of patient conditions may introduce additional challenges that must be fully captured in the retrospective analysis. Therefore, it is important to conduct further studies to evaluate the feasibility and effectiveness of the distant supervision approach in real-time clinical practice and assess its impact on patient outcomes and clinical decision-making processes.

Conclusion

The results of this research provide compelling evidence for the effectiveness of using distant supervision to predict Acute Lung Injury (ALI) in intensive care unit patients. This innovative method demonstrates promising performance in accurately identifying patients at risk of developing ALI within a 24-hour timeframe. These findings contribute to the growing body of evidence supporting the potential of machine learning algorithms in the early detection and prediction of ALI. Identifying patients at risk of ALI before the onset of symptoms or physiological changes is paramount in critical care. It provides a crucial window for intervention, enabling clinicians to initiate appropriate treatments promptly. Timely interventions may help prevent the progression of ALI to more severe conditions, such as Acute Respiratory Distress Syndrome (ARDS), which is associated with higher mortality rates.

The distant supervision-based approach, utilizing machine learning algorithms, offers a novel and effective tool for detecting and predicting ALI. The models demonstrate high accuracy, sensitivity, specificity, and AUC-ROC values, underscoring their potential in clinical practice. These models present an opportunity for healthcare professionals to proactively identify patients at risk, allowing for timely interventions and tailored management strategies to mitigate the progression and complications of ALI.

While the results of this study are promising, further investigations and validation studies are necessary to assess these models' generalizability and clinical applicability across diverse patient populations and healthcare settings. Ensuring these models perform consistently and effectively in real-world scenarios before widespread adoption in clinical practice is essential. By leveraging the vast amounts of data available in electronic medical record datasets, such as MIMIC-III, these models can augment clinical decision-making and support healthcare providers in delivering proactive and personalized care to at-risk patients. Moreover, by enabling the identification of high-risk individuals and the implementation of preventive measures, these models can significantly improve patient outcomes and reduce the burden of ALI-related morbidity and mortality.

Acknowledgments

I want to thank my teachers at the Nightingale-Bamford School for their ongoing support. I would like to thank my research mentor, who has supported me in understanding health informatics and ALI. I would also like to thank the MIT Raise Initiative and SureStart for allowing me to gain experience in creating machine learning models and rendering datasets.

References

1. Matthay, M.A.; Zemans, R.L.; Zimmerman, G.A. Acute respiratory distress syndrome. *Nat. Rev. Dis. Primers* **2019**, *5*(1), 1-22.
2. Calvert, J. S.; Price, D. A.; Chettipally, U. K.; Barton, C. W.; Feldman, M. D. A computational approach to early sepsis detection. *Comput. Math. Methods Med*. **2016**, 2016.
3. Mintz, M.; Bills, S.; Snow, R. Distant supervision for relation extraction without labeled data. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AF NLP: Volume 2.*
4. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; ...; Kudlur, M. TensorFlow: A system for large-scale machine learning. *pp*. 265-283.
5. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.H.; Feng, M.; Ghassemi, M.; et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016,** *3*, 160035. doi: 10.1038/sdata.2016.35.
6. Mikkelsen, M.E.; Shull, W.H.; Biester, R.C.; et al. Cognitive, Mood, and Quality of Life Impairments in a Select Population of ARDS Survivors. *Respirology* **2009**, *14*(1), 76-82. doi: 10.1111/j.1440-1843.2008.01443.x
7. Laffey, J.G.; Madotto, F.; Bellani, G.; et al. Geo-economic variations in epidemiology, patterns of care, and outcomes in patients with acute respiratory distress syndrome: insights from the LUNG SAFE prospective cohort study. *Lancet Respir. Med*. **2017**, *5*(8), 627-638. doi: 10.1016/S2213-2600(17)30213-8
8. Kang, M.G.; Kim, K.H.; Park, S. Early Prediction of Acute Kidney Injury Using Machine Learning Techniques: A Systematic Review of Methodological Efficacy and Clinical Validity. *J. Clin*.
9. Armbrust, M.; Das, T.; Davidson, A.; Ghodsi, A.; Stoica, I. Scaling Spark in the real world: performance and usability. *Proc. VLDB Endow.* **2015**, *8*(12), 1840-1843.
10. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; ...; Kudlur, M. TensorFlow: A system for large-scale machine learning. *pp. 265-283*.
11. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, Vol. 2; Springer, 2009.
12. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **2006**, *27*(8), 861-874.
13. Google Cloud Platform. *Compute Engine*. [Online]. Available: https://cloud.google.com/compute/

DOI: 10.36838/v5i6.8

Author

Eana Shah is a senior at the Nightingale-Bamford School in New York, New York. She plans on majoring in data science and computational biology in the future.