AITA
ARTIFICIAL INTELLIGENCE
THEORY AND APPLICATIONS

■ RESEARCH ARTICLE

Comparing the Performance of Ensemble Methods in Predicting Emergency Department Admissions Using Machine Learning Techniques

Murat Emre Yapıcı[a,†], **Abdulkadir Hızıroğlu**[a], **Ali Mert Erdoğan**[a]

[a] Department of Management Information Systems, İzmir Bakırçay University, İzmir, Türkiye,
[†] m.emreyapici@gmail.com, corresponding author

RECEIVED OCTOBER 30, 2023
ACCEPTED JANUARY 12, 2024

CITATION Yapıcı, M. E., Hızıroğlu, A., & Erdoğan. A. M. (2024). Comparing the performance of ensemble methods in predicting emergency department admissions using machine learning techniques. *Artificial Intelligence Theory and Applications, 4*(1), 11-21.

Abstract

Healthcare data collection, storage, retrieval, and analysis are enabled by various technologies and tools in health information systems. These systems include health information exchanges, telemedicine platforms, clinical decision support systems, and electronic health records. They aim to improve patient outcomes, provider communication, and healthcare workflows. Machine learning is being used in emergency rooms to address challenges such as increasing patient volume, limited resources, and the need for quick decisions. Machine learning algorithms can assist in triage and risk stratification by identifying patients requiring urgent care and predicting the severity of their condition. By analyzing various patient data sources, machine learning can detect patterns and indicators that human clinicians may miss, enabling early intervention and potentially saving lives. However, there is a lack of comparative evaluation of ensemble methods used in analysis. Therefore, this study aims to thoroughly examine and analyze various ensemble methods to understand their efficacy and performance, contributing valuable insights to researchers and practitioners.

Keywords: ensemble methods, logistic regression, prediction, emergency department

1. Introduction

Emergency services are essential healthcare units that provide immediate medical assistance to patients in need. They are categorized based on the urgency and severity of the patient's condition, with red indicating life-threatening emergencies, yellow indicating conditions with a risk of permanent damage, and green indicating mild injuries or illnesses [1]. Information systems play a crucial role in emergency care by providing insights into the workload, patient information, and preliminary assessments in the emergency department. These systems enable informed decision-making for triage and resource allocation, addressing challenges such as overcrowding and improving overall emergency care [2]. Healthcare information systems encompass various technologies, processes, and tools that facilitate the collection, storage, retrieval, and analysis of healthcare data [3]. Electronic health records (EHRs) serve as digital databases of patient information, supporting comprehensive and coordinated care [4]. EHRs aid clinic allergies ending by providing immediate access to vital patient data, alerting healthcare

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2024 İzmir Bakırçay University

professionals to potential interactions or allergies and suggesting evidence-based treatment options [5]. Clinical decision support systems (CDSS) utilize advanced algorithms and medical knowledge databases to enhance diagnosis accuracy, reduce errors, and improve patient safety [5]. Machine learning is a significant component of information systems, analyzing large volumes of medical data and extracting valuable insights. It improves patient care, optimizes resource allocation, and enhances decision-making processes [8]. In emergency departments (EDs), machine learning algorithms improve triage and risk stratification by accurately predicting the severity of a patient's condition and identifying those in urgent need of care [9]. Furthermore, machine learning can detect patterns and indicators in diverse data sources, enabling early diagnosis and prediction of adverse events that may be missed by human clinicians [10]. This study aims to analyze ED admission rates and develop a predictive model to determine the likelihood of future hospitalization. The objectives include reducing overcrowding, expediting treatment for urgent cases, and increasing employee motivation. By analyzing patient demographics, medical history, and severity of conditions, advanced statistical techniques and machine learning algorithms will be used to develop a reliable framework for predicting hospitalization rates. Implementing the study's findings can improve operational efficiency, patient outcomes, and the work environment in EDs.

2. Literature Review

Table 1. Literature Review

STUDY	TECHNIQUE	EVALUATION	RESULT
(Barak-Corren et al., 2021)	**eXtreme Gradient Boosting**	AUC	AUC **0.90-0.93**
(Lee et al. ,2020)	**Multinomial Logistic Regression,** Neural Network, Support Vector Machine	Accuracy (%95 CI)	Accuracy **MLR = 81.6** Neural Network =81.2 Support Vector Machine=81.4
(Graham et al. ,2018)	Logistic Regression, Decision Trees, **Gradient Boosting**	Accuracy (%95 CI) AUC	Accuracy LR=79.94 Decision Trees=80.06 **GBM=80.31** AUC LR=0.849 Decision Trees=0.824 **GBM = 0.859**
(Peck et al. ,2013)	Logistic Regression	AUC R2	AUC **LR=0.80-0.89** R2 LR=0.58 - 0.90
(Woo Suk Hong et al.,2018)	Logistic Regression, **Gradient Boosting,** **Deep Neural Networks**	AUC	AUC **LR=0.87** **XGBOOST=0.87** **DNN=0.87**
(Sun et al., 2011)	Logistic Regression	ROC Accuracy (%95 CI)	ROC **LR=0.849** Accuracy LR=84.7

Ensemble methods in machine learning have garnered significant attention and demonstrated impressive success rates in various applications [11]-[24],[31]. These techniques aim to improve the performance and robustness of predictive models by combining the predictions of multiple base learners, thereby leveraging the diversity of these learners to achieve better overall results. Despite their widespread adoption and promising outcomes, the existing literature still lacks comprehensive comparative studies that thoroughly evaluate and compare different ensemble methods.

Numerous individual studies in the existing literature showcase the effectiveness of ensemble methods, highlighting their contributions to various tasks. For instance, researchers have demonstrated the benefits of ensemble methods like XGBoost Regression in classification tasks. A study by Barak-Corren et al. (2010) [11] showed that an ensemble of Multinomial Logistic Regression models outperformed individual logistic regression models in predicting customer churn, achieving higher accuracy and better generalization.

However, despite these individual success stories, there is a notable lack of direct comparisons between different ensemble methods in the literature. Few studies conduct head-to-head evaluations to determine which ensemble technique is more suitable for specific scenarios.

For this reason, this study aims to fill this gap by comparing the performance of four prominent ensemble methods: Adaboost, LogitBoost, GentleBoost, and RusBoost.

3. Research Methodology

3.1. Dataset

The dataset used in this study consists of 1267 systematically selected records of adult patients admitted to two emergency departments between October 2016 and September 2017 [25], [32] . Emergency Service. In order to ensure accurate forecasting, certain columns in the dataset needed to be removed, which could potentially impact the accuracy of the predictions. Hence, it is important to highlight the current state of the dataset as it undergoes estimation. The resulting configuration of the dataset is presented below.

These parameters are key indicators used in medical assessments. The mental scale assesses a person's level of consciousness and responsiveness, ranging from alertness to unconsciousness. The Numeric Rating Scale (NRS) for pain measures pain intensity on a numerical scale. Systolic blood pressure (SBP) represents the pressure in arteries when the heart beats, while diastolic blood pressure (DBP) is the pressure when the heart rests between beats. Respiration rate (RR) measures the number of breaths per minute, crucial in evaluating respiratory health. Saturation indicates the oxygen saturation level in the blood, often measured with a pulse oximeter, reflecting the amount of oxygen carried by red blood cells. Together, these parameters provide a comprehensive snapshot of a person's mental state, pain level, cardiovascular health, respiratory function, and oxygenation status, aiding healthcare professionals in making informed decisions about treatment and care.

3.2. Data Preparation

Tasks such as data cleaning, integration, transformation (min, max on all features), and feature selection are involved in this process. One important modification made during data preparation was transforming disposition values into binary categories. This simplification enables easier interpretation and analysis of the dataset, specifically regarding patient outcomes (discharged or admitted).

The Emergency Department categorized and analyzed patients based on variables such as group, sex, age, arrival mode, injury, mental status, and pain. This approach provided insights into patient cohorts, gender patterns, age trends, arrival modes, types of injuries,

mental well-being, and discomfort levels. Considering these dimensions allowed for a deeper understanding of the patient population and facilitated in-depth data analysis.

Table 2. Descriptions and Distributions of the variables

Variable	Descriptions and Distributions of the variables
Sex	1: Female (51.8%) / 2: Male (48.2%)
Age	Age (mean: 53.9, std: 18.8)
Patients_number_per_hour	Patients number/hours (mean: 7.5, std: 3.1)
Arrival_mode	1: Walking (6.6%) / 2: 119 use (19.5%) / 3: Private car(61.8%) / 4: Private ambulance (10.7%) / 5: Others (1.4%)
Injury	1: Non-injury (80.9%) / 2: Injury (%19.1)
Mental	1: Alert (95.4%) / 2: Verbal response (2.4%) / 3: Pain response (1.7%) / 4: Unconsciousness (0.5%)
Pain	1: Pain (56.9%) / 2: Non-pain (43.1%)
NRS_pain	Numeric rating scales of pain (between 1-5 (84.2%) / between 6-10(15.8%))
SBP	Systolid blood pressure (mean: 131.6, std: 26.7)
DBP	Diastolic blood pressure (mean: 78.6, std: 14.6)
HR	Heart rate (mean: 82.2, std: 16.4)
RR	Respiration rate (mean: 19.3, std: 1.9)
BT	Body temperature (mean: 36.3, std: 0.7)
Saturation	Saturation to use pulse oximeter (mean: 96.9, std: 4.2)
Disposition	0: Discharge (68.1%) / 1: Admission (31.9%)

To enhance analysis integrity and reliability, missing values and outliers were handled. By identifying and eliminating these problematic data points, the analysis was strengthened in terms of robustness and accuracy. Categorical variables were encoded using One-hot Encoding, representing each unique value as a separate column. This streamlined the dataset and uncovered patterns and correlations. The dataset was divided 80/20 into training and test samples. The training sample was used for model training, while the test sample assessed their performance, ensuring reliable and suitable analysis methods.

3.3. Modelling and Evaluation

Logistic Regression is a popular choice for binary classification tasks due to its simplicity, interpretability, and proven effectiveness. It estimates the probability of an event occurring based on input variables, making it reliable in various domains [26]. To enhance the predictive performance, ensemble methods such as AdaBoost [27], LogitBoost [28], GentleBoost [29], and RUSBoost [30] were employed. These methods combine multiple models to improve accuracy and handle class imbalance. AdaBoost iteratively trains weak classifiers, focusing on misclassified samples, while LogitBoost optimizes Logistic Regression parameters. GentleBoost assigns smaller weights to misclassified samples to reduce sensitivity to outliers, and RusBoost addresses class imbalance by under sampling the majority class. Since Logistic Regression is used very frequently in this field, we added it to the benchmarking study. Lastly, performance metrics such as Sensitivity, Specificity and F1 Score and so on , which are frequently used in Binary Classification, were used to evaluate and compare model performances. The following provides a brief explanation of performance metrics.

Sensitivity: Ratio of true positive (discharged) examples correctly predicted by the model. It is calculated using the formula:

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

Specificity: Ratio of true negative (admitted) examples correctly predicted by the model. It is calculated using the formula:

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

Precision: Ratio of true positive examples correctly predicted by the model. It is calculated using the formula:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Negative Predictive Value: Ratio of true negative examples correctly predicted by the model. It is calculated using the formula:

$$Negative\ Predictive\ Value = \frac{TN}{TN + FN} \quad (4)$$

False Positive Rate: Ratio of true negatives incorrectly predicted as positives. It is calculated using the formula:

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \quad (5)$$

False Discovery Rate: Ratio of positives predicted incorrectly as positives. It is calculated using the formula:

$$False\ Discovery\ Rate = \frac{FP}{FP + TP} \quad (6)$$

False Negative Rate: Ratio of true positives incorrectly predicted as negatives. It is calculated using the formula:

$$False\ Negative\ Rate = \frac{FN}{TP + FN} \quad (7)$$

Accuracy: Ratio of correct predictions (both true positive and true negative) to the total number of examples. It is calculated using the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

F1 Score: Harmonic mean of precision and sensitivity, providing a balance between the two metrics. It is calculated using the formula:

$$f_1\ Score = 2 * \frac{P * R}{P + R} \tag{9}$$

Matthews Correlation Coefficient: Calculates the correlation coefficient between observed and predicted binary classifications. It takes values between -1 and +1, where +1 indicates perfect predictions, 0 implies no improvement over random guessing, and -1 signifies complete disagreement between prediction and observation.

Area Under the Receiver Operating Characteristic Curve (AUC): Quantifies the performance of a binary classification model across various threshold values. The ROC curve illustrates the relationship between true positive rate and false positive rate for different threshold values. AUC represents the area under this curve.

4. Results

ROC RESULTS

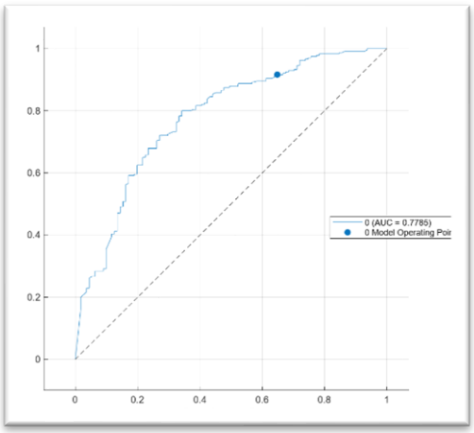

Figure 1. AdaBoost

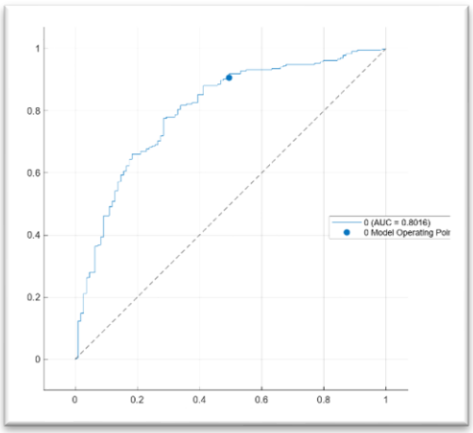

Figure 2. Logistic Reg.

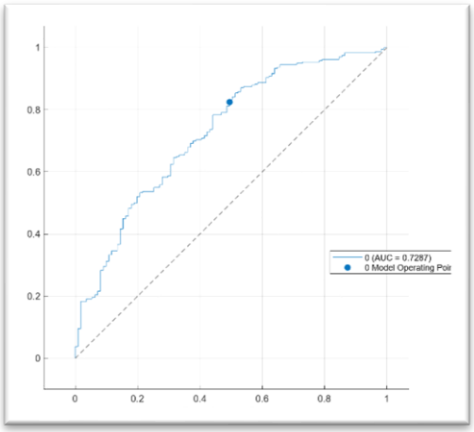

Figure 3. Gentleboost

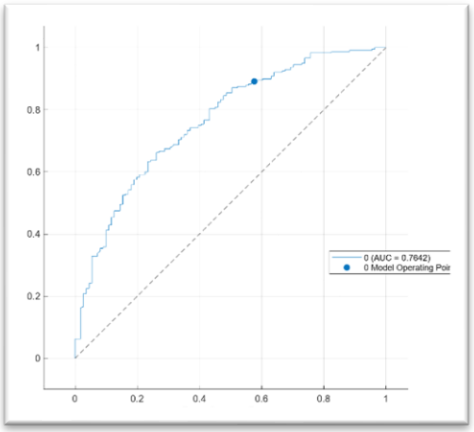

Figure 4. Logitboost

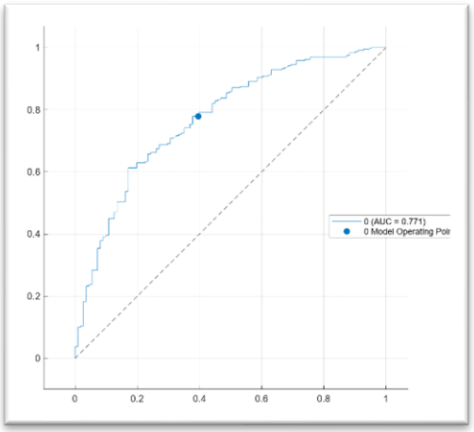

Figure 5. Rusboost

Table 3. Results and Comparison of Methods

	Sensitivity	Specificity	Precision	Negative Predictive Value	False Positive Rate	False Discovery Rate	False Negative Rate	Accuracy	F1 Score	Matthews Correlation Coefficient	AUC
Adaboost	0.753	0.661	0.916	0.351	0.339	0.083	0.246	0.737	0.827	0.333	0.778
Logistic Reg.	0.795	0.717	0.908	0.500	0.282	0.091	0.204	0.778	0.848	0.457	0.801
Gentleboost	0.782	0.571	0.825	0.504	0.428	0.175	0.217	0.723	0.803	0.341	0.728
Logitboost	0.769	0.643	0.891	0.423	0.356	0.108	0.230	0.743	0.826	0.361	0.764
Rusboost	0.809	0.558	0.779	0.603	0.441	0.220	0.190	0.723	0.794	0.375	0.771

The Adaboost model performs reasonably well but has room for improvement. It shows good sensitivity (0.7534) but comparatively lower specificity (0.6610), indicating challenges in accurately identifying negative cases. The model has high precision (0.9167) but a relatively low negative predictive value (0.3514), suggesting a considerable number of incorrect negative predictions. The false positive rate (0.3390) is moderately high, while the false discovery rate (0.0833) is low. The false negative rate (0.2466) can be improved for better performance. The model's accuracy is 0.7379, and the F1 score (0.8271) demonstrates a reasonable balance between precision and sensitivity. The Matthews Correlation Coefficient (0.3333) indicates moderate overall agreement. AUC value of 0.778 indicates that the model has good discrimination ability, distinguishing between classes with moderate accuracy. In summary, while the Adaboost model has some positive aspects, improvements can be made in terms of specificity, negative predictive value, false positive rate, false negative rate, and overall accuracy through optimization and fine-tuning.

The Logistic Regression model performs well with good sensitivity (0.7956) and specificity (0.7179). It has high precision (0.9083) and relatively low false positive rate (0.2821) and false discovery rate (0.0917). The negative predictive value (0.5000) can be improved, indicating room for better identification of negative cases. The false negative rate (0.2044) is relatively low. The model's accuracy is 0.7784, and the F1 score (0.8482) demonstrates a good balance between precision and sensitivity. The Matthews Correlation Coefficient (0.4579) indicates moderate overall agreement. With a AUC score of 0.801, LR has the best result comparing the other models, which means a higher level of accuracy in classifying outcomes based on the model's predictions. In summary, Logistic Regression model shows reliable performance with high sensitivity, specificity, precision, and accuracy. Improvements can be made in the negative predictive value and false negative rate through fine-tuning and optimization efforts.

The Gentleboost model shows mixed results with potential for improvement. It has a sensitivity of 0.7826, correctly identifying a decent proportion of positive cases. However, it struggles with specificity (0.5714) in accurately identifying negative cases. The precision (0.8250) is relatively high, with a majority of positive predictions being correct. The negative predictive value (0.5045) suggests room for improvement in correctly identifying negative cases. The false positive rate (0.4286) is relatively high, indicating a considerable number of negative cases being falsely classified as positive. The false discovery rate (0.1750) is relatively low, suggesting fewer false positive predictions. The false negative rate (0.2174) represents the proportion of positive cases incorrectly classified as negative, which can be further improved. The model's accuracy is 0.7236, and the F1 score (0.8032) demonstrates a reasonable balance between precision and sensitivity. The Matthews Correlation Coefficient (0.3416) indicates a moderate level of overall agreement. Meanwhile, AUC value of 0.728 suggests fair discrimination ability, with some limitations in accurately separating classes. In summary, the Gentleboost model shows a mix of strengths and weaknesses, with room for improvement in specificity, negative predictive value, and false positive rate. Further optimization and fine-tuning efforts are needed to enhance its performance.

The Logitboost model shows reasonable performance. It has a sensitivity of 0.7698, correctly identifying a decent proportion of positive cases, and a specificity of 0.6438, indicating reasonable performance in identifying negative cases. The precision (0.8917) is relatively high, with a majority of positive predictions being correct. However, the negative predictive value (0.4234) suggests room for improvement in correctly identifying negative cases. The false positive rate (0.3562) is moderately high, implying some negative cases being falsely classified as positive. The false discovery rate (0.1083) is

relatively low, indicating fewer false positive predictions. The false negative rate (0.2302) represents the proportion of positive cases incorrectly classified as negative, which is moderate. The model's accuracy is 0.7436, and the F1 score (0.8263) demonstrates a reasonable balance between precision and sensitivity. The Matthews Correlation Coefficient (0.3610) indicates a moderate level of overall agreement. AUC score of 0.764 indicates moderately good performance in distinguishing between classes. In summary, the Logitboost model shows moderate performance with good precision and sensitivity. However, improvements can be made in terms of specificity, negative predictive value, false positive rate, and overall accuracy. Further optimization and fine-tuning efforts may enhance its performance.

The Rusboost model shows mixed performance. It has a sensitivity of 0.8095, correctly identifying a relatively high proportion of positive cases, but struggles with specificity (0.5583) in accurately identifying negative cases. The precision (0.7792) is moderate, with a majority of positive predictions being correct. The negative predictive value (0.6036) is relatively high, indicating better performance in correctly identifying negative cases. The false positive rate (0.4417) is relatively high, implying a substantial number of negative cases being falsely classified as positive. The false discovery rate (0.2208) is moderately high, indicating a significant number of false positive predictions. The false negative rate (0.1905) represents the proportion of positive cases incorrectly classified as negative, which is relatively low but can be improved. The model's accuracy is 0.7236, and the F1 score (0.7941) demonstrates a reasonable balance between precision and sensitivity. The Matthews Correlation Coefficient (0.3752) indicates a moderate level of overall agreement. AUC value of 0.771 denotes decent discriminatory power, although slightly lower compared to the other models scores, especially LR, but still indicating a reasonable level of predictive accuracy. In summary, the Rusboost model shows mixed performance with strengths in sensitivity and negative predictive value, but weaknesses in specificity and false positive rate. Further optimization and fine-tuning may be necessary to enhance its overall performance.

Rusboost is a rarely encountered ensemble method that provides a comparative perspective. It outperforms Gentleboost and closely resembles Adaboost, suggesting it as a valuable alternative with similar predictive accuracy. The inclusion of Rusboost expands the knowledge base and promotes exploration of ensemble methodologies. AUC analysis reveals remarkable similarity between Rusboost and Adaboost, showcasing favorable outcomes. This warrants a reevaluation of common approaches and encourages further research on Rusboost's capabilities. Its success enhances analytical outcomes and expands possibilities for future studies.

5. Conclusion

Healthcare data analysis relies on various technologies and systems, including health information exchanges, telemedicine platforms, clinical decision support systems, and electronic health records. Machine learning has revolutionized emergency care by improving triage and risk stratification. Machine learning algorithms accurately identify patients needing urgent care and predict the severity of their conditions, enabling early intervention. Despite a wide range of analysis methods, there is a lack of comparative evaluations of ensemble methods. This study aims to comprehensively examine and analyze ensemble methods for healthcare data analysis. Logistic Regression consistently performs the best, followed closely by Adaboost. Rusboost, an underutilized method, shows promising performance similar to Adaboost. Logitboost also demonstrates comparable results. Gentleboost, however, is the least successful method. These findings highlight the importance of careful selection of ensemble methods for

specific prediction studies in the Emergency Department. Researchers can make informed decisions to advance predictive models in emergency care.

Additionally, there are several limitations of the study. One notable limitation is related to the dataset used for analysis, where the availability and quality of the data might introduce inherent biases and limitations in representing the full spectrum of emergency care cases. Moreover, the selection of ensemble methods for analysis might limit the comprehensiveness of the comparison, as other relevant techniques not included could impact the overall conclusions. Furthermore, the study's generalizability might be constrained by the specific context and settings in which the research was conducted, considering different healthcare systems, patient populations, or emergency care protocols.

In future, researchers can carefully select a diverse and representative dataset of emergency care cases, considering different medical conditions, patient demographics, and severity levels. To address privacy concerns, they can work with de-identified or synthetic datasets to ensure compliance with regulations while maintaining the dataset's integrity. Given potential limitations in implementing certain ensemble methods in the healthcare context, the researchers can adopt a focused approach by comparing a subset of ensemble methods that are more suitable for the specific emergency care prediction task. This targeted comparison can ensure the study's relevance and feasibility within the given constraints. To assess the performance of the ensemble methods accurately, researchers can choose appropriate performance metrics aligned with the specific goals of emergency care prediction. Considering the restricted availability of healthcare data, the researchers will utilize techniques like cross-validation and bootstrapping to obtain more reliable estimates of ensemble method performance. These resampling methods will enable them to evaluate the ensemble methods on multiple subsets of the data, yielding more robust and generalizable results.

References

[1] Medicana (n.d.). Tıbbi Servisler ve Acil servis. https://www.medicana.com.tr/tibbi-birimler/acil-servis

[2] Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency medicine*, 52(2), 126-136.

[3]] Mamlin, B. W., Biondich, P. G., Wolfe, B. A., Fraser, H., Jazayeri, D., Allen, C., ... & Tierney, W. M. (2006). Cooking up an open source EMR for developing countries: OpenMRS–a recipe for successful collaboration. *In AMIA Annual Symposium Proceedings (Vol. 2006, p. 529).* American Medical Informatics Association.

[4] McGinn, C. A., Grenier, S., Duplantie, J., Shaw, N., Sicotte, C., Mathieu, L., ... & Gagnon, M. P. (2011). Comparison of user groups' perspectives of barriers and facilitators to implementing electronic health records: a systematic review. BMC medicine, 9(1), 1-10.

[5] Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj, 330(7494), 765.

[6] Agency for Healthcare Research and Quality (n.d.). Clinical Decision Support. https://www.ahrq.gov/cpi/about/otherwebsites/clinical-decision-support/index.html#:~:text=Clinical%20decision%20support%20(CDS)%20provides,team%20and%20patient%20to%20consider

[7] HealthIT.gov (n.d.). What is an electronic health record (EHR)?. https://www.healthit.gov/faq/what-electronic-health-record-ehr#:~:text=EHRs%20are%20a%20vital%20part,decisions%20about%20a%20patient's%20care

[8] Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. Critical care medicine, 46(4), 547.

[9] Kunt, M. M. (2021). Emergency Medicine and Artificial Intelligence. https://dergipark.org.tr/en/download/article-file/1985451

[10] Sun, J., Zhang, Y., & Tang, L. (2019). Predicting patient deterioration in the emergency department: A machine learning approach. *Journal of Biomedical Informatics*, 98, 103267.

[11] Barak-Corren, Y., Chaudhari, P., Perniciaro, J., Waltzman, M., Fine, A. M., & Reis, B. Y. (2021). Prediction across healthcare settings: a case study in predicting emergency department disposition. *npj Digital Medicine,* 4(1), 169.

[12] Lee, S. H., Chinnam, R. B., Dalkiran, E., Krupp, S., & Nauss, M. (2020). Prediction of emergency department patient disposition decision for proactive resource allocation for admission. *Health Care Management Science*, *23*(3), 339–359. https://doi.org/10.1007/s10729-019-09496-y

[13] Chen, C., Hsieh, J., Cheng, S., Lin, Y., Lin, P., & Jeng, J. (2020). Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. *International Journal of Medical Informatics*, *139*, 104146. https://doi.org/10.1016/j.ijmedinf.2020.104146

[14] LaMantia, M. A., Platts-Mills, T. F., Biese, K., Khandelwal, C., Forbach, C. R., Cairns, C. B., Busby-Whitehead, J., & Kizer, J. S. (2010). Predicting Hospital Admission and Returns to the Emergency Department for Elderly Patients. *Academic Emergency Medicine*, *17*(3), 252–259. https://doi.org/10.1111/j.1553-2712.2009.00675.x

[15] Parker, C. A., Liu, N., Wu, S. X., Shen, Y., Lam, S. S. W., & Ong, M. E. H. (2019). Predicting hospital admission at the emergency department triage: A novel prediction model. *American Journal of Emergency Medicine*, *37*(8), 1498–1504. https://doi.org/10.1016/j.ajem.2018.10.060

[16] Graham, B., Bond, R., Quinn, M., & Mulvenna, M. (2018). Using data mining to predict hospital admissions from the emergency department. *IEEE Access*, *6*, 10458-10469.

[17] Zhang, X., Kim, J., Patzer, R. E., Pitts, S. R., Patzer, A., & Schrager, J. D. (2017). Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods of information in medicine*, *56*(05), 377-389.

[18] Peck, J. S., Gaehde, S. A., Nightingale, D. J., Gelman, D. Y., Huckins, D. S., Lemons, M. F., ... & Benneyan, J. C. (2013). Generalizability of a simple approach for predicting hospital admission from an emergency department. Academic Emergency Medicine, 20(11), 1156-1163.

[19] Sun, Y., Heng, B. H., Tay, S. Y., & Seow, E. (2011). Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine*, *18*(8), 844-850.

[20] Roquette, B. P., Nagano, H., Marujo, E. C., & Maiorano, A. C. (2020). Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Networks*, *126*, 170-177.

[21] Mowbray, F., Zargoush, M., Jones, A., de Wit, K., & Costa, A. (2020). Predicting hospital admission for older emergency department patients: Insights from machine learning. *International Journal of Medical Informatics*, *140*, 104163.

[22] Peck, J. S., Benneyan, J. C., Nightingale, D. J., & Gaehde, S. A. (2012). Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, *19*(9), E1045-E1054.

[23] Leegon, J., Jones, I., Lanaghan, K., & Aronsky, D. (2005). Predicting hospital admission for Emergency Department patients using a Bayesian network. In *AMIA Annual Symposium Proceedings* (Vol. 2005, p. 1022). American Medical Informatics Association.

[24] Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PloS one*, *13*(7), e0201016.

[25] Kaggle (2021, August 28). *Emergency Service - Triage Application*. https://www.kaggle.com/datasets/ilkeryildiz/emergency-service-triage-application

[26] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[27] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

[28] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[29] Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In *icml* (Vol. 99, pp. 124-133).

[30] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE transactions on systems, man, and cybernetics-part A: systems and humans*, *40*(1), 185-197.

[31] Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PloS one*, *13*(7), e0201016.

[32] Moon, S. H., Shim, J. L., Park, K. S., & Park, C. S. (2019). Triage accuracy and causes of mistriage using the Korean Triage and Acuity Scale. *PloS one*, *14*(9), e0216972.