American Journal of Emergency Medicine 84 (2024) 141–148

Contents lists available at ScienceDirect

American Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/ajem

Development and validation of a machine learning framework for improved resource allocation in the emergency department

Abdel Badih el Ariss [a], Norawit Kijpaisalratana [a], Saadh Ahmed [b], Jeffrey Yuan [c], Adriana Coleska [a], Andrew Marshall [d], Andrew D. Luo [a,d], Shuhan He [a,*]

[a] Emergency Department, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America
[b] Georgia State University, Department of computer science, Atlanta, Georgia
[c] Northwestern University, Department of Data science, Evanston, IL, United States of America
[d] Emergency Department, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States of America

ARTICLE INFO

Article history:
Received 15 April 2024
Received in revised form 3 July 2024
Accepted 24 July 2024

Keywords:
Artificial intelligence
Machine learning
Triage
Emergency department operations

ABSTRACT

Objective: The Emergency Severity Index (ESI) is the most commonly used system in over 70% of all U.S. emergency departments (ED) that uses predicted resource utilization as a means to triage [1], Mistriage, which includes both undertriage and overtriage has been a persistent issue, affecting 32.2% of total ED visits [2]. Our goal is to develop a machine learning framework that predicts patients' resource needs, thereby improving resource allocation during triage.

Methods: This retrospective study analyzed ED visits from the Medical Information Mart for Intensive Care IV, dividing the data into training (80%) and testing (20%) cohorts. We utilized data available during triage, including patient vital signs, age, gender, mode of arrival, medication history, and chief complaint. Azure AutoML was used to create different machine learning models trained to predict the 144 target columns including laboratory panels and imaging modalities as well as medications required during patients' ED visits. The 144 models' performance was evaluated using the area under the receiver operating characteristic curve (AUROC), F1 score, accuracy, precision and recall.

Results: A total of 391,472 ED visits were analyzed. 144 Voting ensemble models were created for each target. All frameworks achieved on average an AUC score of 0.82 and accuracy of 0.76. We gathered the feature importance for each target and observed that 'chief complaint', among others, had a high aggregate feature importance across different targets.

Conclusion: This study shows the high accuracy in predicting resource needs for patients in the ED using a machine learning model. This can greatly improve patient flow and resource allocation in already resource limited emergency departments.

1. Introduction

In the high-stakes environment of emergency departments (ED), the triage process plays a pivotal role. This process is not just about directing patient flow to appropriate treatment areas, but it ensures effective allocation of resources for care delivery. While multiple different triage systems exist, the Emergency Severity Index (ESI) is the most commonly used in the US, employed in over 70% of all U.S. EDs [1]. The ESI is a five-level triage algorithm that categorizes patients into groups from 1 (most urgent) to 5 (least urgent) on the basis of their vital signs and resource needs. The ESI has been widely adopted in EDs worldwide due to its reliability and validity in predicting hospital admission and resource utilization [2,3].

Despite advancements in triage methods, significant challenges persist, particularly with the Emergency Severity Index (ESI), leading to frequent mistriage. While these tools rely on objective data for triage decisions, the inherent subjective judgment of clinicians during triage plays a critical role, introducing variability and potential bias [4,5]. Mistriage is defined here as the act of giving a patient an inappropriate level of triage, including both undertriage and overtriage. Undertriage refers to a situation where a patient who need urgent interventions is mistakenly assigned a lower ESI level (for example ESI 3 or 4 instead of a 2). On the other hand, overtriage occurs when a patient who requires minimal resources is assigned a higher ESI level (for example ESI 3 or 4 instead of a 5). Mistriage has been found to occur in 32.2% of total ED visits, of which 3.3% were undertriaged and 28.9% were

* Corresponding author.
E-mail address: she@mgh.harvard.edu (S. He).

https://doi.org/10.1016/j.ajem.2024.07.040
0735-6757/

overtriaged [4,5]. This issue stems from the subjective nature of the ESI system, particularly for ESI levels 3, 4, and 5, which require clinicians to estimate the number of resources patients are expected to need. For example, a patient with mild abdominal pain might be overtriaged to ESI level 3 if the clinician anticipates the need for multiple resources such as tests and consultations, whereas they might more appropriately be assigned to ESI level 4 or 5 if fewer resources are needed. Conversely, a patient with moderate symptoms might be undertriaged to ESI level 4 but may actually require an ESI level 3 due to needing several diagnostic tests and treatments. Incorrectly estimating resource needs can lead to assigning an inappropriate triage level, potentially delaying necessary treatment and straining ED resources.

This highlights the need to improve the triage process by providing more data-driven tools for resource allocation decisions, which will in turn inform the assignment of a patient's ESI score.

An additional aspect of this endeavor is the early decision-making regarding patient disposition, especially in cases requiring transfers. The ability to determine the need for specialized care or transfer to another facility can be vital for patient outcomes and efficient use of ED resources [6,7]. Artificial intelligence (AI) and machine learning (ML) emerge as potent tools to transcend these challenges, capable of sifting through complex data to uncover patterns and enhance decision-making thereby supporting more informed and timely decisions in emergency care. In addition, they can mitigate the tendency to overuse diagnostic tools by providing data-driven suggestions [8]. Studies have already shown the effectiveness of AI in various scenarios including mortality, admission or need for critical care [9,10].

Furthermore, the utilization of Electronic Health Records (EHR) in conjunction with AI algorithms has led to the development of predictive models that can anticipate outcomes like hospitalization needs and critical care events. These models, which encompass machine learning, deep learning, and interpretable machine learning, are promising but require further comparative studies to ensure their reliability and generalizability [10].

Our research aims to leverage machine learning models to accurately predict patients' resource needs, thus improving the efficacy of the Emergency Severity Index (ESI) system. This includes predicting the need for laboratory tests, imaging services, medications, and procedures which are significant contributors to resource utilization and length of stay in the ED. Accurate prediction of resource needs enables us to refine the ESI system, leading potentially to better patient allocation across various ED sections. This improvement not only enhances efficiency and reduces wait times but also significantly boosts patient outcomes, particularly for those who are currently misclassified.

2. Methods

2.1. Study design, setting, participant selection

This retrospective cohort study analyzed clinical data from patients aged 18 and over who visited the ED of Beth Israel Deaconess Medical Center, an academic medical center in Boston, Massachusetts, between 2008 and 2019. The data, sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, included 425,087 ED visits [11]. Since the MIMIC-IV database does not involve identifiable human subjects, Institutional Review Board (IRB) approval was not required for this study.

2.2. Data pre-processing

The dataset used for all algorithms was sourced from triage information, comprising both structured and unstructured data elements. Structured data included vital signs such as temperature, blood pressure (systolic and diastolic), heart rate, respiratory rate, and oxygen saturation, as well as demographic details such as age, sex, and mode of arrival. Additionally, a categorized list of each patient's medications, organized by therapeutic purpose, was incorporated. Unstructured data, such as the free-text descriptions of patients' chief complaints, was also included. This type of text allows clinicians to capture the exact symptoms and nuances reported by patients, which can be crucial for accurate diagnosis and treatment [12]. Furthermore, there is no universal coding for chief complaints, which could limit the generalizability of the model [13].

To conduct the machine learning analysis, a dataset was created with a schema consisting of feature columns to serve as inputs and target columns to serve as outputs.

Encounters with no recorded vital signs were removed from the dataset, and the missing values in the remaining encounters were imputed using mean imputation. The feature columns included structured data such as age and pulse rate were added with their raw values, categorical data like medication used at home were one-hot encoded which converts them into a binary format suitable for machine learning applications. It achieves this by assigning a unique column in a matrix to each medication category. Within this matrix, the presence of a medication is indicated by a 1, while its absence is marked with a 0, thereby transforming the categorical information into a format that machine learning algorithms can efficiently process. On the other hand, unstructured data like chief complaints was manipulated by azure Auto Machine Learning (AutoML) using different techniques like word embeddings, TF-IDF vectorization or categorical hash among others suited for machine learning depending on the model used.

The target columns encompass various resources utilized, including laboratory testing, imaging techniques, procedures, and medications administered. Laboratory tests were grouped into detailed panels to assess specific organ systems. For instance, amylase and lipase were collectively categorized under pancreatic enzymes. Imaging techniques were classified into distinct categories such as MRI, CT, ultrasound, and X-ray. Procedures (such as interventional radiology) were consolidated to indicate whether a patient underwent any procedure. Lastly, the medication data in the ED focused on the top 90% most frequently used medications, excluding the least used 10% due to their minimal count, which rendered them statistically insignificant for predictive analysis. These columns were converted into one-hot encoded format, using 0 s and 1 s to denote whether they were used or not.

The dataset underwent a stratified sampling into an 80% training and a 20% testing. Furthermore, to address the issue of imbalanced data for each target due to the sheer number of samples, we employed a random-balanced under-sampling technique to the training set. This method balances the class distribution in the dataset by reducing the size of the more abundant class. Such balancing ensures that the model can learn effectively from both classes, thereby enhancing its generalization capabilities and improving its accuracy in making predictions across all classes, not just the predominant one.

The dataset was uploaded to Azure AutoML for preprocessing and feature engineering, which automated the transformation of raw data into a machine learning-ready format. During this phase, AutoML undertook the imputation of missing values by substituting them with the feature's mean, ensuring that no data points were discarded due to incomplete information.

2.3. Model development

After preprocessing the initial dataset, it was then split into several datasets, each corresponding to a single target column. This resulted in different subsets, each with their own unique training dataset. These datasets were then uploaded to Azure Blob storage in a tabulated format with each column data type predefined. Following this, an experiment was created for each prediction target with its dataset specified in the configuration. Azure AutoML was then triggered to execute training for each experiment.

Following the initial preprocessing, Azure AutoML proceeds to the model selection and training phase, specifically within the context of a classification task. Utilizing its extensive repository of algorithms,

Table 1
Demographic and clinical characteristics of training and testing sets.

Demographic	Training Set (N = 313,177)	Testing Set (N = 78,295)	p-value
Male, n (%)	141,761 (45.27)	35,947 (45.91)	<0.001
Age, mean (SD)	50.00 (20.00)	50.04 (20.03)	0.62
Race			
White, n (%)	168,296 (53.74)	42,300 (54.03)	0.07
Black, n (%)	57,406 (18.33)	14,377 (18.36)	37
Acuity Level, n (%)			
1	11,323 (3.62)	2889 (3.68)	0.21
2	104,933 (33.51)	26,179 (33.43)	0.33
3	174,023 (55.57))	43,483 (55.53)	0.42
4	22,081 (7.05)	5530 (7.1)	0.35
5	817 (0.25)	214 (0.26)	0.31
Mode of Arrival, n (%)			
Walk-in	193,100 (61.66)	48,205 (61.57)	0.32
Ambulance	110,186 (35.18)	27,569 (35.21)	0.43
Vital Signs			
Heart Rate (beats/min), mean (SD)	84.97 (17.76)	84.50 (17.50)	<0.001
Systolic Blood Pressure (mmHg), mean (SD)	135.12 (41.56)	134.93 (22.43)	0.08
Diastolic Blood Pressure (mmHg), mean (SD)	81.50 (11.89)	79.58 (17.81)	<0.001
Body Temperature (F), mean (SD)	98.02 (4.03)	98.00 (2.56)	0.09
Respiratory Rate (breaths/min), mean (SD)	17.57 (9.36)	17.53 (4.02)	0.07
Oxygen Saturation (percents), mean (SD)	98.51 (19.04)	98.43 (5.08)	0.04

AutoML selects a diverse array of classification models that range from traditional methods, such as logistic regression and decision trees, to more complex ones like gradient boosting, SVM and LightGBM. It then applies these models within a robust validation framework, employing a 5-fold cross-validation technique to rigorously evaluate model performance. Cross-validation is a resampling procedure used to assess the predictive capabilities of a model by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In each of the five iterations, a different subset of the data is held out as the test set while the remaining data is used for training, ensuring that every data point is used for both training and validation exactly once. The performance metrics from these folds are then aggregated to provide a comprehensive measure of model efficacy. Upon completion of the model training and validation process, Azure AutoML synthesizes the insights gathered from various individual classifiers to construct a superior predictive model.

This is achieved through the implementation of a voting ensemble method among 12 trained machine learning models, which leverages soft voting. In soft voting, predictions from each constituent model are weighted and combined based on the predicted probabilities of the class labels, rather than a simple majority vote. This probabilistic approach allows for a more nuanced aggregation of model predictions, giving higher weight to the confidence level of each model's predictions. The resulting ensemble model capitalizes on the strengths of its members while mitigating individual weaknesses, culminating in a robust and often more accurate final model.

2.4. Model performance evaluation

To provide a comprehensive assessment of model performance on the test set, several key metrics were utilized. After retrieving the best performing models, the voting ensemble runs again on the test set and retrieves the F1 score, AUC, accuracy, precision score, and recall score. The Accuracy metric measures the proportion of correct predictions, including both true positives and true negatives, among the total number of cases examined, serving as a straightforward and intuitive gauge of overall model performance. The AUC (Area Under the Curve) reflects the model's ability to discriminate between positive and negative classes, with values closer to 1 indicating better discrimination. The F1 Score combines precision and recall into a single metric, offering a balance between the model's precision in predicting positive instances and its ability to recall all positive instances. Precision Score focuses exclusively on the model's accuracy in predicting positive instances, whereas the Recall Score measures the model's ability to identify all actual positives [14].

3. Results

A total of 391,472 ED visits were included in this analysis, with 313,177 visits in the training set and 78,295 visits in the testing set (Table 1). The input features used in model development are shown in Table 2, with each variable in our study database having <6% missing values. We created 144 voting ensemble models for each target.

3.1. Statistical analysis

3.1.1. Performance Metrics of the voting ensembles of several resources

Fig. 1 displays the performance metrics of voting ensemble classifiers used in medical settings to predict various patient needs, including hospital admission, imaging and procedure requirements, laboratory testing, and medication use in the ED. The model predicting hospital admissions shows promising results, with an accuracy of approximately 0.75 and an AUC close to 0.8, indicating strong overall performance and discriminative power. The F1 score, at 0.7, signifies an excellent balance between precision and recall, confirming the model's reliability. Precision at nearly 0.65suggests that the positive predictions are correct, while a recall around 0.7 underscores the model's ability to identify most actual cases that require admission.

The model's capability in forecasting the need for imaging and procedures is similarly proficient in general accuracy (0.78) and identifying cases (AUC of 0.85). However, the lower F1 (0.15) and precision (0.09)

Table 2
Patient features and the percentage of missing values.

Feature	Data type	Missing values, n (%)
Gender	Categorical	0 (0%)
Age	Continuous	0 (0%)
Race	Categorical	0 (0%)
Mode of Arrival	Categorical	0 (0%)
Chief Complaint	Text	0 (0%)
Systolic Blood Pressure	Continuous	18,291 (4.03%)
Diastolic Blood Pressure	Continuous	19,091 (4.49%)
Body Temperature	Continuous	23,415 (5.51%)
Heart Rate	Continuous	17,090 (4.02%)
Respiratory Rate	Continuous	20,353 (4.79%)
Oxygen Saturation	Continuous	20,596 (4.85%)
Home Medications	Categorical	0 (0%)

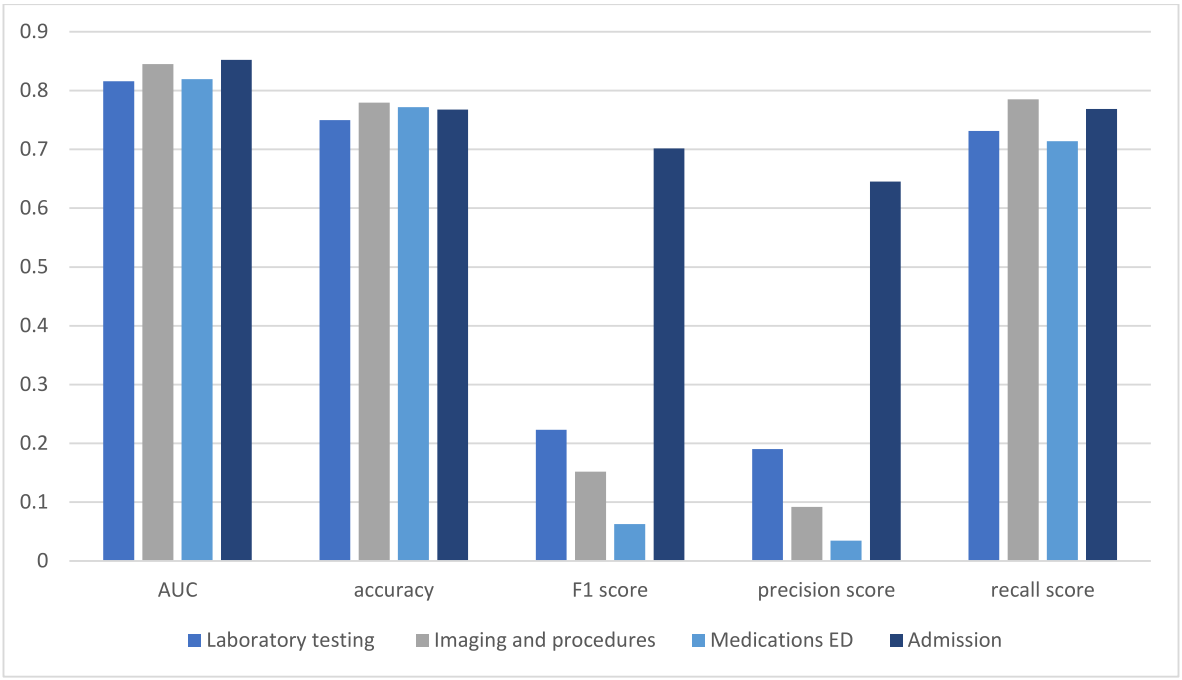

Fig. 1. Average Performance of the voting ensembles that predicts the laboratory testing, imaging and procedures, Medications used in the ED and admission.

scores point to a significant area for enhancement, especially in minimizing false positives. For laboratory tests, the classifier's accuracy (0.75) and AUC (0.81) imply effective discrimination between necessary and unnecessary tests. Nevertheless, the lower F1 score (0.22) and precision score (0.2) reveal a propensity for over-predicting positive cases. In predicting medication use in the ED, the classifier mirrors the overall high accuracy (0.77) and AUC (0.81) seen in the performance of other models. However, the model's low F1 score (0.06) and precision score (0.03) highlight a skewed emphasis on recall over precision.

3.2. Closer examination of some voting ensembles

Upon closer examination of the predictive models, we find a noteworthy that the model predicting hospital admissions assigns the greatest importance to chief complaints (as illustrated in Fig. 2). Among the chief complaints, Fig. 3 illustrates some of the words in the chief complaints that were associated with a higher likelihood of admission.

Further analysis reveals that age is the second most influential factor, with admission rate increasing with age, as shown in Fig. 4. Additionally, blood pressure readings significantly impact the likelihood of admission; patients with extremely low (50–100) or high (200–250) readings are more likely to be flagged for admission (Fig. 5).

Focusing on imaging classification, the model predicting the utilization of CT scans assigns high feature importance to chief complaints such as “alcohol status post fall” at 1.75 and “right lower quadrant abdominal pain” at 1.57, indicating these as strong predictors for the need of a CT scan, as illustrated in Fig. 6.

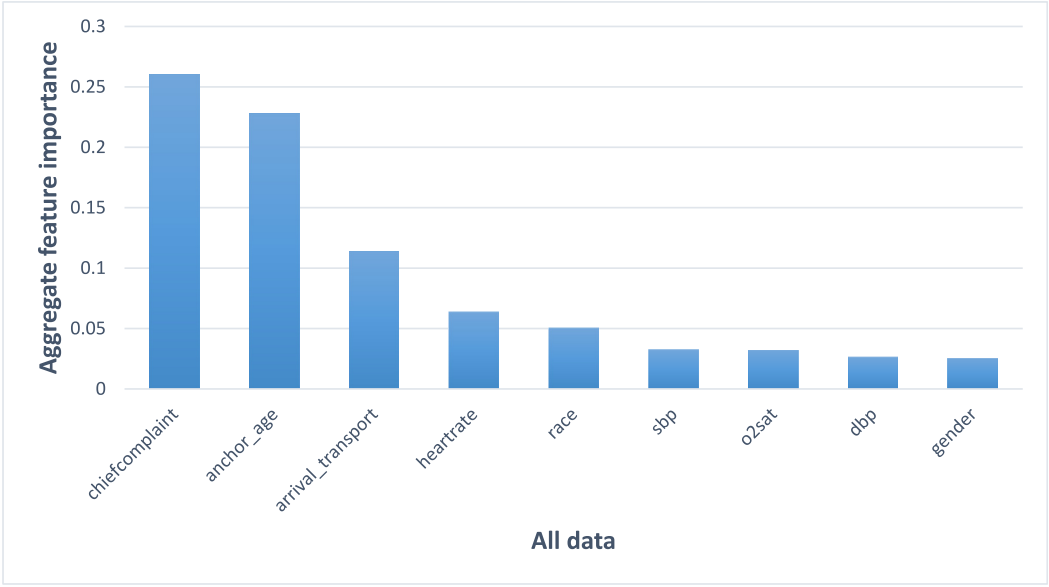

Fig. 2. Aggregate Feature importance for admission.

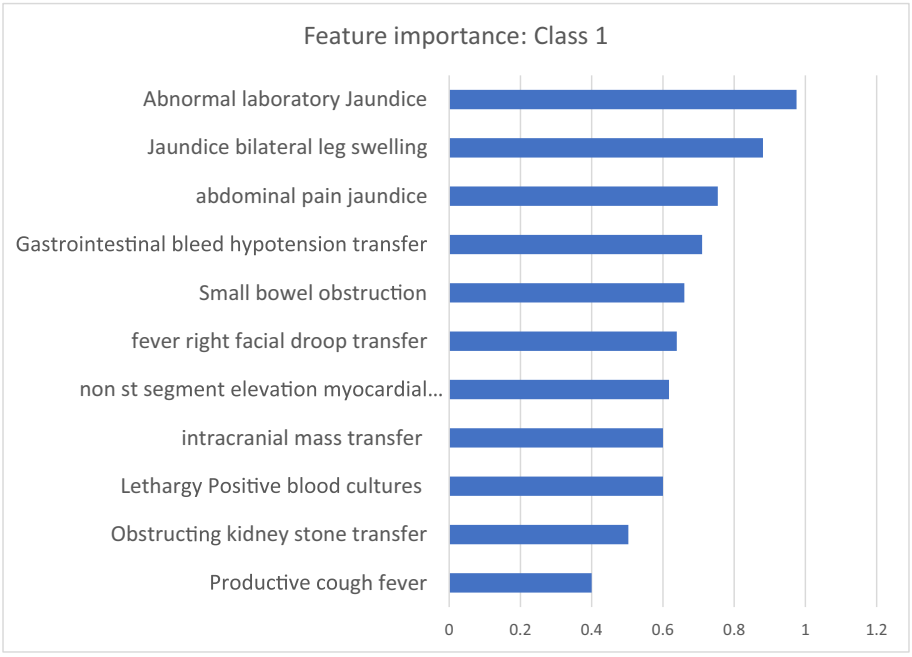

Fig. 3. Plot showing the detailed feature importance of the chief complaint for admission.

In laboratory testing, the machine learning model for pleural fluid metabolic panel (i.e. PH, proBNP, LDH, protein etc) accurately identifies patients with O2 saturation below 95% as needing pleural fluid analysis, typically associated with pleural effusions (Fig. 7).

To illustrate the rationale behind our model's classification process, we chose "Emtricitabine/Tenofovir" from the list of predicted medications as an example. This model identifies patients prescribed home antiretroviral as indicator for administering Emtricitabine Tenofovir in the ED with a feature importance reaching 2.4 (Fig. S8).

Highlighting the capabilities of other medication prediction models, a prime illustration is the model developed for forecasting the necessity of norepinephrine. This model pinpoints individuals whose systolic

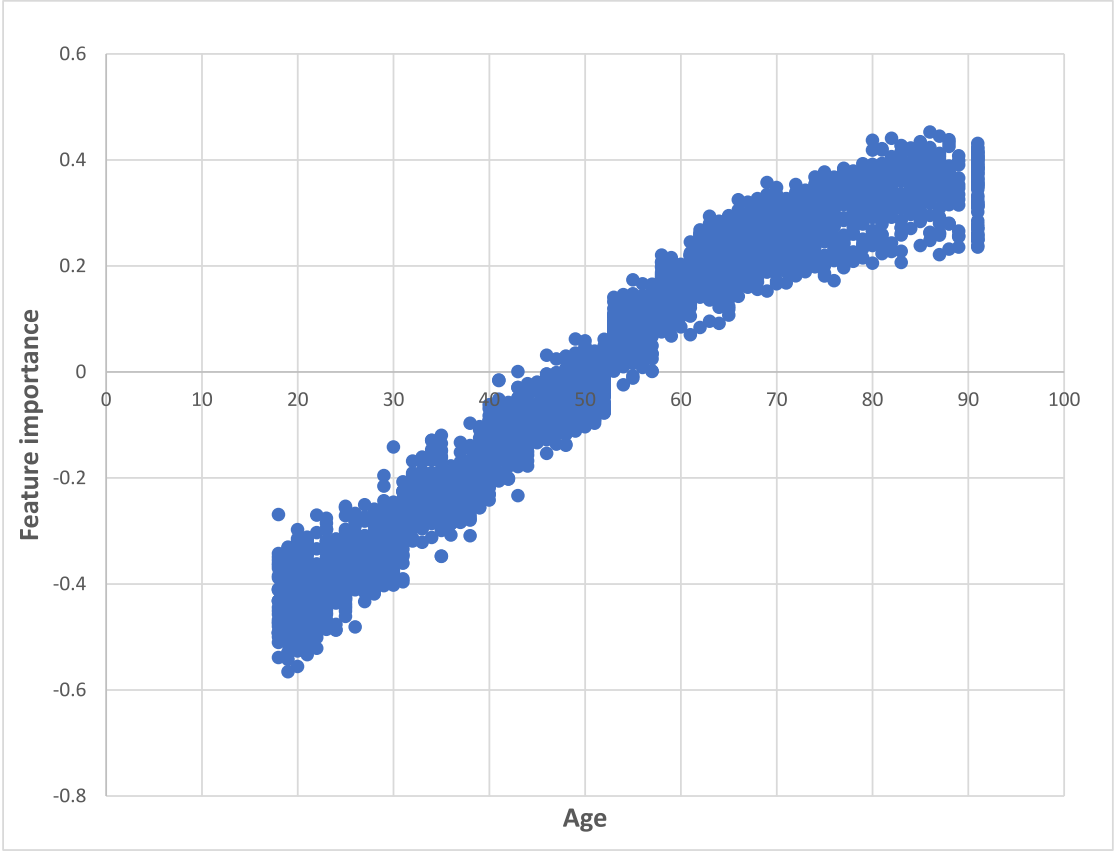

Fig. 4. Detailed Analysis of age as a Key Feature in admitting a patient as indicated by the voting ensemble.

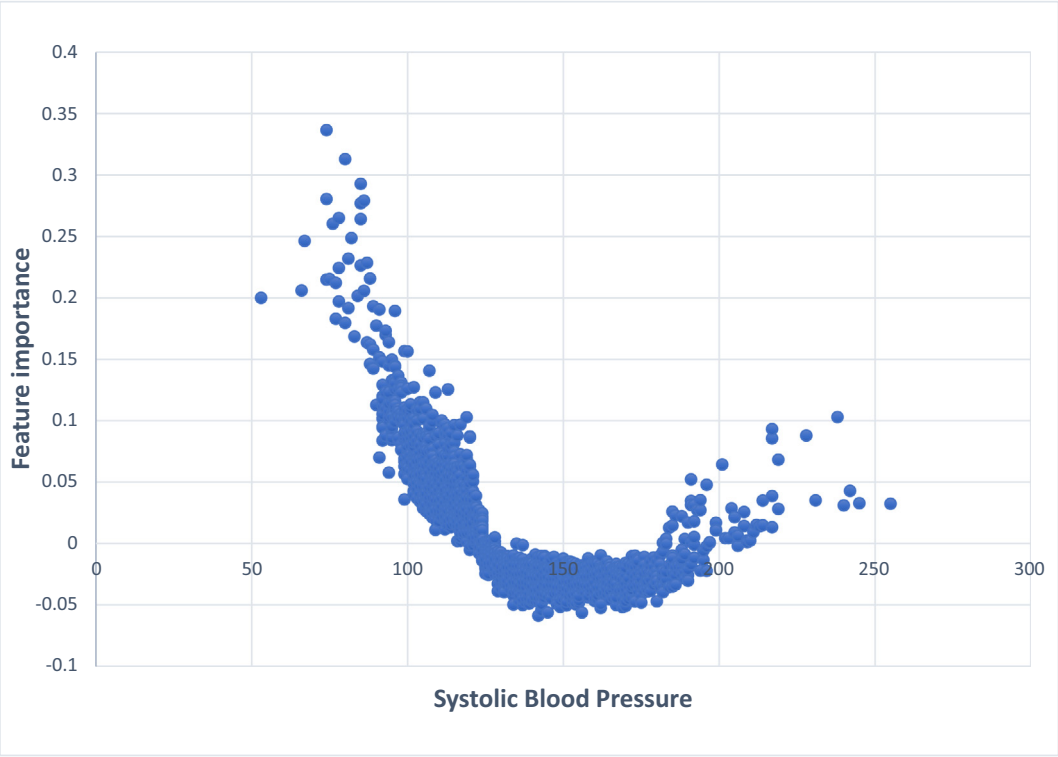

Fig. 5. Detailed Analysis of systolic blood pressure (sbp) as a Key Feature in admitting a patient as indicated by the voting ensemble.

blood pressure (SBP) falls below 100 mmHg as likely candidates in need of norepinephrine (Fig. S9).

4. Discussion

In our study, we developed 144 machine learning models to anticipate the need for various resources in the ED, using data from patient triage. These models, informed by historical data on resource allocation, learn to recognize patterns and make objective predictions about resource needs. Our model's performance is comparable to previous work in this domain, such as the study by Hunter-Zinck et al., which used multilabel machine learning framework to predict clinical orders simultaneously [15]. In comparison, our study employed a different approach by using a separate model for each target instead of a multilabel classification. This method provided us with clearer insights into how the models classify the necessity for each order based on the triage information and thus the clinical profile of the patient. However, both studies reached the same conclusion that machine learning holds significant potential for predicting resources and supporting decision-making in emergency departments.This study's evaluation of the performance of voting ensembles in predicting medication usage in ED settings reveals a multifaceted picture of the different model's efficacy (Figs. 1). Notably, the high accuracy observed indicates a general reliability of the models in making correct predictions about imaging, laboratory

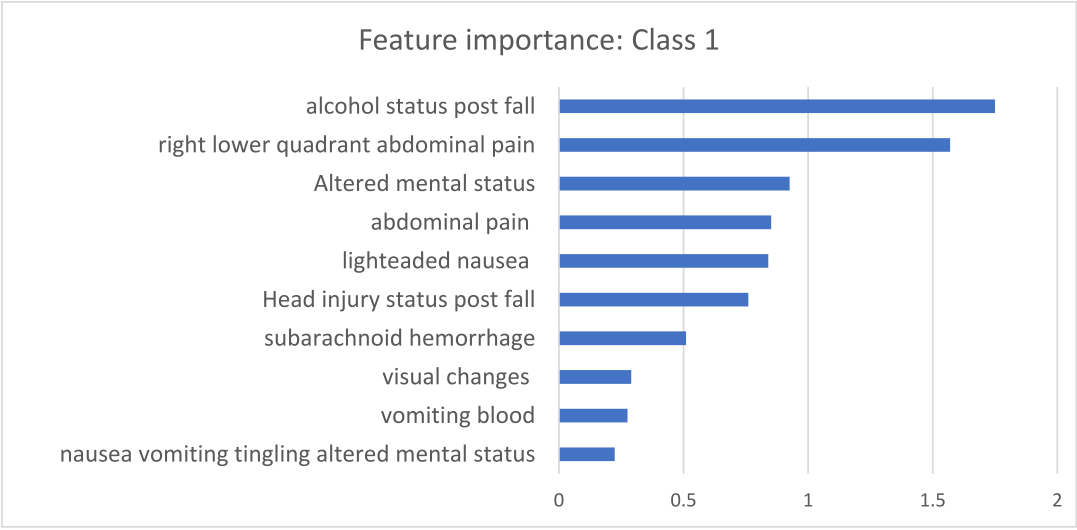

Fig. 6. Detailed Analysis of chief complaint as a Key Feature in performing a Computed tomography (CT) as indicated by the voting ensemble.

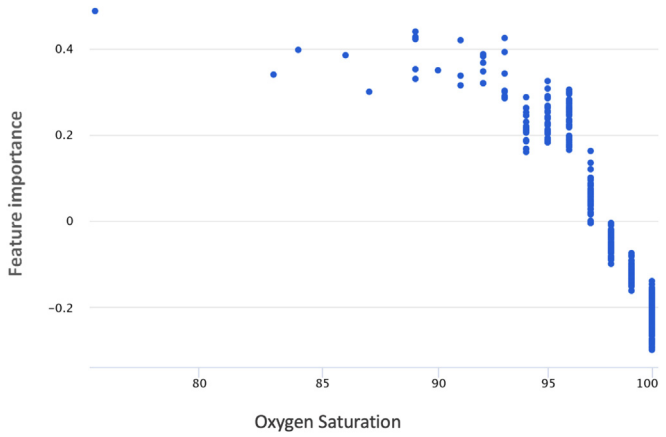

Fig. 7. Detailed Analysis of O2sat as a Key Feature in performing a pleural fluid analysis as indicated by the voting ensemble.

tests and medication to be ordered in the ED. Such a level of accuracy is reassuring, given the critical nature of decision-making in emergency medicine. Complementing this, the high AUC scores demonstrate the model's robust capability to discriminate effectively between instances where medication is required and where it is not. This characteristic is particularly vital in the ED context, where the accurate identification of medication necessity can significantly impact patient outcomes.

However, the study also uncovers areas needing improvement, particularly regarding precision metrics. The lower precision score indicates a reduced confidence in the model's positive predictive power. This reduced precision could lead to challenges in distinguishing true positives from false positives, raising concerns about the potential for over-prescription or unnecessary medication administration in the ED. The lower F1 score, representing the harmonic mean of precision and recall, points to an imbalance in the models' performance. The tendency of the models to favor recall over precision suggests a proclivity towards identifying cases that require medication at the expense of generating more false positives. In practical terms, this could result in the recommendation of medications for patients who do not need them, leading to unwarranted treatment interventions. Lastly, the precision score further underscores the issue of the models mistakenly labeling negative instances (cases where no medication is needed) as positive.

While the models demonstrate strong predictive accuracy and recall, thus effectively identifying patients in need of a medication, the lower precision-related metrics highlight a crucial area for enhancement. However, it is important to emphasize that the models are trained based on patterns observed in the behaviors of emergency physicians in the ED, particularly regarding their resource ordering practices, which is concordant with the literature [16].

As we investigate and explore the mechanism through which the voting ensemble model predicts admissions, we find that it bases its decisions on various 'chief complaints,' as well as 'age' (Fig. S2–5) aligning with current literature [17]. Moreover, as blood pressure readings below 100 are typically regarded as an unstable condition, Fig. S4 shows that the model is effectively predicting hospital admissions for this specific patient group. Furthermore, Fig. S9 underscores the model's proficiency in accurately recommending the use of norepinephrine for patients with low blood pressure, particularly in cases where blood pressure falls below 90. This demonstrates the model's effectiveness in critical clinical decision-making scenarios. In addition, an examination of the voting ensemble model predicting the need for CT scans, as illustrated in Fig. S6, reveals that 'chief complaint' ranks highly in terms of feature importance influencing its decision-making. Notably, specific complaints like 'alcohol status post-fall', 'right lower quadrant abdominal pain', and 'altered mental status' are among the key reasons indicating the necessity for a CT scan. On the other hand, the model also identifies scenarios where a CT is not required, such as in patients presenting with finger laceration and other cases. This distinction aligns with clinical expectations, as conditions like 'alcohol status post-fall' is often associated with various bodily injuries that could necessitate a CT scan [18].

Fig. S8 shows that patients who take at home an antiretroviral are more likely to get prescribed "Emtricitabine Tenofovir" in the ED; which is an antiretroviral especially used in HIV patients [19]. This underlines the performance of one of the models dedicated to predict antiretroviral administration in the ED.

In summary, a comprehensive evaluation of selected voting ensembles reveals their reliability in classifying the necessity of resources. This detailed analysis confirms the models' capability in making dependable decisions about resource allocation in clinical settings. It's important to emphasize that these models do not rely on a single parameter to predict the use of specific resources. Instead, they incorporate a comprehensive view of the patient's context, including vital signs, demographics, medication history, and chief complaint. This holistic approach allows for a more accurate and tailored prediction of resource utilization.

Enhancing the model's accuracy and predictive power could be achieved by incorporating a greater volume and variety of data points, particularly those gathered post-admission. Such additional data would enable the training of the model on a broader range of patient outcomes, thereby improving its predictive capabilities for those who are admitted to the hospital.

Issues such as sociodemographic biases and tendencies like over-triaging to an ESI of level 2 or under-triaging older patients [20,21] suggest the necessity for a more data-driven approach. Our machine learning models, by analyzing patient data at admission, can effectively predict patient needs, thereby substantially minimizing variability in the ESI's framework [22].

By integrating these models into the ESI process, we can significantly enhance the system's objectivity. Leveraging objective data derived from patterns identified by machine learning models, the ESI will become more robust, thus optimizing patient flow and empowering clinicians with more informed decision-making during triage [23]. Moreover, the potential for automation in triage with these models adds another layer of efficiency. Finally, the dynamic nature of machine learning models means they continuously learn and improve with new data. As they encounter more patient cases, they refine their predictive accuracy, adapting to evolving clinical environments.

4.1. Limitations

Our study has certain limitations, one of which is using mean imputation to replace missing data. This method can reduce the variance of variables and affect their relationships [24]. Additionally, it can introduce bias, especially if the data is not missing at random, leading to inaccurate estimates and predictions [25].

In future work, we plan to explore more advanced approaches, such as regression or deep learning imputation, to address these issues. In addition, the absence of certain essential resources such as consultations with specialists, EKG and use of monitors which were not documented in the database. Furthermore, the observed low F1 and precision scores of the voting ensemble suggest a reduced efficiency in recognizing false positives, potentially due to the characteristics of the training dataset. Additionally, the models were trained solely on data from a tertiary academic medical center, potentially limiting their generalizability to similar hospitals. To address this, we plan to incorporate data from multiple institutions in future work, enhancing the model's accuracy and applicability across diverse healthcare settings.

5. Conclusions

This study acts as a demonstrative framework, showcasing how data can be utilized by a suite of machine learning models to predict the entire spectrum of resources a patient might require in the ED. This data-driven approach aims to eliminate bias and minimize error, forecasting potential admissions and standardizing the triage process. Further research is planned to broaden the scope of these models, incorporating more comprehensive data sets and increasing the number of models.

Meeting

SAEM conference 2024, Phoenix, Arizona**,** NERDS conference.

Funding sources/disclosures

None, AA, NK, SA, JY, AM, AL and SH reports no conflict of interest.

Data sharing statement

The machine learning models can be shared upon request.

Author contribution statement

Abdel Badih El Ariss, Norawit Kijpaisalratana, and Shuhan He came up with the study idea. Abdel Badih El Ariss, Norawit Kijpaisalratana, and Saadh Ahmed built the machine learning models. All authors, including Abdel Badih El Ariss, Norawit Kijpaisalratana, Saadh Ahmed, Adriana Coleska, Andrew Marshall, Andrew D. Luo, and Shuhan He, helped write and improve the manuscript.

CRediT authorship contribution statement

Abdel Badih el Ariss: Writing – review & editing, Writing – original draft, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Norawit Kijpaisalratana:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Saadh Ahmed:** Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Jeffrey Yuan:** Writing – original draft. **Adriana Coleska:** Writing – original draft. **Andrew Marshall:** Writing – original draft. **Andrew D. Luo:** Writing – original draft. **Shuhan He:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

None.

Acknowledgments

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ajem.2024.07.040.

References

[1] Mccugh M, Tanabe P, Mcclelland M, Khare RK. More patients are triaged using the emergency severity index than any other triage acuity system in the United States. Acad Emerg Med. 2012;19(1):106–9. https://doi.org/10.1111/j.1553-2712.2011.01240.x.

[2] Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC. The emergency severity index triage algorithm version 2 is reliable and valid. Acad Emerg Med. 2003;10:1070–80.

[3] Chmielewski, Nicholas DNP, RN, CEN, Cenp, NEA-BC, Faen Moretz, Jason MHA, BSN, RN, CEN, CTRN. ESI Triage Distribution in U.S. Emergency Departments. Adv. Emerg. Nurs. J. 44(1):p 46–53, January/March 2022. https://doi.org/10.1097/TME.0000000000000390

[4] Sax DR, Warton EM, Mark DG, et al. Assessment of emergency severity index triage accuracy and disparities in a large, diverse cohort. JAMA Netw Open. 2023;6(1):e222222. https://doi.org/10.1001/jamanetworkopen.2022.22222.

[5] Smith M, Cattermole G, Li X. Evaluation of the emergency severity index in US emergency departments for the rate of Mistriage. JAMA Netw Open. 2023;6(1):e2802556. https://doi.org/10.1001/jamanetworkopen.2022.02556.

[6] Schreiber M, Yin S, O'Neil B, Moore C, Mello MJ. Using machine learning to predict patient transfer in the emergency department. J Am Med Inform Assoc. 2018;25(3):271–7. https://doi.org/10.1093/jamia/ocx115.

[7] Wu CC, Yen ZS, Wu MH. Artificial intelligence and machine learning in emergency medicine. Emerg Med Clin North Am. 2020;38(1):153–63. https://doi.org/10.1016/j.emc.2019.10.001.

[8] Pines JM, Hollander JE, Localio AR, Metlay JP. The association between physician risk tolerance and imaging use in abdominal pain. Am J Emerg Med. 2009;27(5):552–7. https://doi.org/10.1016/j.ajem.2008.04.014.

[9] Tan L, Young J, Ong MEH, et al. A systematic review of the use of machine learning in the prediction of the emergency severity index triage level. J Healthc Eng. 2021;2021:6645260. https://doi.org/10.1155/2021/6645260.

[10] Clifton DA, Clifton L, Sandu DM, et al. Machine learning for early prediction of hospitalization in the emergency department. J Am Med Inform Assoc. 2019;26(10):1–7. https://doi.org/10.1093/jamia/ocz112.

[11] Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset [published correction appears in Sci Data. 2023 Jan 16;10(1):31. https://doi.org/10.1038/s41597-023-01945-2] [published correction appears in Sci Data. 2023 Apr 18;10(1):219. https://doi.org/10.1038/s41597-023-02136-9]. *Sci Data*. 2023;10(1):1. Published 2023 Jan 3. https://doi.org/10.1038/s41597-022-01899-xJohnson A., Bulgarelli L., Pollard T., Celi L.A., Mark R., Horng S. MIMIC-IV-ED (version 2.2). PhysioNet 2023. https://doi.org/10.13026/5ntk-km72.

[12] Chang David, Hong Woo Suk. Richard Andrew Taylor, Generating contextual embeddings for emergency department chief complaints. JAMIA Open. 2020;Volume 3(2):160–6. https://doi.org/10.1093/jamiaopen/ooaa022.

[13] Horng S, Greenbaum NR, Nathanson LA, McClay JC, Goss FR, Nielson JA. Consensus development of a modern ontology of emergency department presenting problems-the hierarchical presenting problem ontology (HaPPy). Appl Clin Inform. 2019 May;10(3):409–20. https://doi.org/10.1055/s-0039-1691842. [Epub 2019 Jun 12. PMID: 31189204; PMCID: PMC6561773].

[14] Microsoft. (n.d.). How to understand automated machine learning. Retrieved from https://learn.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml?view=azureml-api-2

[15] Hunter-Zinck Aley S, Peck Jordan S, Strout Tania D, Gaehde Stephan A. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. J Am Med Inform Assoc. December 2019;26(12):1427–36. https://doi.org/10.1093/jamia/ocz171.

[16] Lam JH, Pickles K, Stanaway FF, et al. Why clinicians overtest: development of a thematic framework. BMC Health Serv Res. 2020;20:1011. https://doi.org/10.1186/s12913-020-05844-9.

[17] Panahpour Eslami N, Nguyen J, Navarro L, et al. Factors associated with low-acuity hospital admissions in a public safety-net setting: a cross-sectional study. BMC Health Serv Res. 2020;20:775. https://doi.org/10.1186/s12913-020-05456-3).

[18] Taylor TR, Mhlanga J, Thomas A. Alcohol-related head injury: impact on acute CT workload in a major trauma center published in the Br. J. Neurosurg. (2009;23(6):622–624. https://doi.org/10.3109/02688690903215666.

[19] Mayo Clinic. (n.d.). Emtricitabine and Tenofovir (Oral Route) Description. Retrieved from https://www.mayoclinic.org/drugs-supplements/emtricitabine-and-tenofovir-oral-route/description/drg-20061833

[20] Sangal RB, Su H, Khidir H, et al. Sociodemographic disparities in queue jumping for emergency department care. JAMA Netw Open. 2023;6(7):e2326338. https://doi.org/10.1001/jamanetworkopen.2023.26338.

[21] Malinovska A, Pitasch L, Geigy N, Nickel CH, Bingisser R. Modification of the emergency severity index improves mortality prediction in older patients. West J Emerg Med. 2019;20(4):633–40. https://doi.org/10.5811/westjem.2019.4.40031.

[22] Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23(1):64. https://doi.org/10.1186/s13054-019-2351-7.

[23] Wardi G, Carlile M, Holder A, Shashikumar S, Hayden SR, Nemati S. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. Ann Emerg Med. 2021;77(4):395–406. https://doi.org/10.1016/j.annemergmed.2020.11.007.

[24] Jakobsen JC, Gluud C, Wetterslev J, et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol. 2017;17:162. https://doi.org/10.1186/s12874-017-0442-1.

[25] Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Can J Cardiol. 2021 Sep;37(9):1322–31. https://doi.org/10.1016/j.cjca.2020.11.010. Epub 2020 Dec 1. PMID: 33276049; PMCID: PMC8499698.