A hybrid machine learning and natural language processing model for early detection of acute coronary syndrome

Joshua Emakhu
*Henry Ford Health System*

Egbe Etu Etu
*San Jose State University, egbe-etu.etu@sjsu.edu*

Leslie Monplaisir
*Wayne State University*

Celestine Aguwa
*Wayne State University*

Suzan Arslanturk
*College of Engineering*

See next page for additional authors

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

**Recommended Citation**

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Authors
Joshua Emakhu, Egbe Etu Etu, Leslie Monplaisir, Celestine Aguwa, Suzan Arslanturk, Sara Masoud, Imokhai Theophilus Tenebe, Hashem Nassereddine, Mohamed Hamam, and Joseph Miller

This article is available at SJSU ScholarWorks: https://scholarworks.sjsu.edu/faculty_rsca/4669
A hybrid machine learning and natural language processing model for early detection of acute coronary syndrome

Joshua Emakhu a, Egbe-Etu Etu b,**, Leslie Monplaisir c, Celestine Aguwa c, Suzan Arslanturk d, Sara Masoud e, Imokhai Theophilus Tenebe e, Hashem Nassereddine f, Mohamed Hamam f, Joseph Miller f

a Department of Public Health Sciences, Henry Ford Health, Detroit, MI, 48202, USA
b Department of Marketing and Business Analytics, San Jose State University, San Jose, CA, 95192, USA
c Department of Industrial & Systems Engineering, Wayne State University, Detroit, MI, 48202, USA
d Department of Computer Science, Wayne State University, Detroit, MI, 48202, USA
e San Jose State University, San Jose, CA, 95192, USA
f Departments of Emergency Medicine, Henry Ford Health, Detroit, MI, 48202, USA

ARTICLE INFO
Handling Editor: Madijd Tavana
Keywords:
Machine learning
Natural language processing
Heart disease
Feature importance
Emergency department

ABSTRACT
Acute coronary syndrome (ACS) is a leading cause of mortality and morbidity. Predicting the associated risks of patients with chest pain using electronic health record data can help identify those needing more tailored care. This study proposes the development of a reliable prediction framework to serve as a diagnostic support tool for preventing misdiagnoses among patients with clinical concerns for ACS. Data were collected from an urban, demographically diverse hospital in Detroit, Michigan, for patients presenting to the emergency department (ED) with primary chief complaints of chest pain from January 2017 to August 2020. This study incorporated term frequency-inverse document frequency features from free-text summaries, which contain anecdotal symptom descriptions and are among the first data points provided upon entering the ED. The analysis included 16,096 patients with clinical concerns for ACS and trained three machine learning models, logistic regression, AdaBoost, and linear discriminant analysis, across different data processing stages to predict patients with ACS from non-ACS etiology. The AdaBoost model outperformed the other two models with an accuracy of 94% and an F1-score of 0.943 in predicting ACS on the testing data. This study identified key independent factors from patient demographics, comorbidities, and clinical narrative data that predicted patients in ACS. The prediction framework can serve as a decision-support tool to classify ACS and inform physicians about better ACS risk factors.

1. Introduction

Between 2015 and 2020, the two leading causes of death were heart disease and cancer, accounting for approximately 43.5% of all deaths during this timeframe [1]. According to a recent report from the American Heart Association, cardiovascular disease claims more lives annually than all types of cancer and chronic lower respiratory diseases [2]. One of the most common types of heart disease in the United States is coronary artery disease, which affects blood flow to the heart [3]. Acute coronary syndrome (ACS) is a major manifestation of this disease. ACS comprises a spectrum of illnesses categorized into three distinct processes, each with its own diagnostic and therapeutic challenges [4]: (i) ST-segment elevation myocardial infarction (STEMI), in which myocardial ischemia results from total coronary artery occlusion; (ii) Non-ST-segment elevation myocardial infarction (NSTEMI), in which myocardial ischemia results from partial occlusion; and (iii) Unstable angina (UA), in which there is dynamic plaque occlusion of the coronary artery, placing patients at high risk for ischemia. Myocardial infarction, commonly known as a heart attack, is the lack of blood flow to the heart muscle, whereas occlusion is the complete or partial blockage of a blood vessel. Appropriate management of ACS and timely interventions are crucial to minimize the mortality and morbidity associated with cardiovascular diseases. Nonetheless, this requires us to understand past trends and patterns of ACS-related mortality and inform policies regarding imminent tailored interventions based on the availability of data and models [5].

** Corresponding author.
E-mail addresses: jemakhu1@hfhs.org (J. Emakhu), egbe-etu.etu@sjsu.edu (E.-E. Etu).

https://doi.org/10.1016/j.health.2023.100249
Received 19 June 2023; Received in revised form 17 August 2023; Accepted 25 August 2023
Available online 1 September 2023
2772-4425/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
The unavailability of timely and precise disease status or factor information can lead to adverse outcomes. Clinical records can be used to improve the availability of timely information, which can enhance the quality of medical care to emphasize correctly diagnosed information. Hence, it is vital to extract and predict correct information from a given clinical record. Currently, clinical records are stored electronically in various flexible formats. An unstructured dataset provides the liberty to record jargon in a clinical record that can be perceived by a specific group of people, which makes the extraction process difficult. The extensive implementation of electronic health record (EHR) systems in medical service organizations permits the collection of rich clinical data from many patients [6]. Extensive EHR data quality as follows: 1) make more precise prediction models considering a broad range of patient characteristics; 2) have the option to refresh these prediction models more frequently with minimal analytical efforts; and 3) enhance the nature of these prediction models with fewer issues, such as the common generalization problem [7].

Nonetheless, EHRs are puzzling to represent and model because of their high dimensionality, noise, heterogeneity, sparseness, incompleteness, arbitrary errors, and systematic biases. Furthermore, an abundance of information about patient clinical history is mostly locked behind free-text clinical notes [8], as composing text remains the most natural and expressive strategy for recording clinical events. The growth of natural language processing (NLP) techniques is fundamental for automatically transforming clinical notes into structured data that can be directly processed using machine learning (ML) algorithms. The utilization of NLP in the clinical domain is increasing with various applications including the identification of biomedical hypotheses from radiology reports [9], nursing documentation [10], and discharge summaries [11]. However, frameworks built on NLP applied to clinical narratives have not been widely used in clinical settings to aid decision support systems or work processes.

The introduction of cutting-edge analytics, such as ML with text mining techniques and algorithms, offers the possibility of a more productive use of unstructured EHR data for clinical research [12–14]. Identifying the clinical subdomains of a clinical note, such as cardiology, gastroenterology, and neurology, might be valuable for enhancing the viability of clinical predictive analytics by considering specialty-associated conditions [15]. Knowledge of the clinical subdomain aids in the ensuing data and knowledge extraction steps. Developing algorithms for the automatic classification of patient attributes, exposures such as smoking, or disease status in characterized categories might be considered for easier entry into real-world data for studies of safety, effectiveness, and treatment patterns of pharmaceutical product use in routine clinical practice [16].

Prior research on automated document classification used rule-based knowledge engineering by physically employing a set of expert intelligence rules [17]. Currently, ML algorithms, such as logistic regression and kernel methods [18,19], and NLP techniques have been utilized to aid clinical decision making through risk stratification [20,21], disease status, or progression prediction using clinical notes. For example, researchers have used ML and NLP to perform automated clinical document classification to adjust intensive care risk through procedure and diagnosis identification [22], detecting cardiovascular failure criteria [23], identifying adverse drug effects [24,25], identifying the status of autism spectrum disorder [19], asthma [26], COVID-19 vaccination response [27], and the activity of rheumatoid arthritis [28]. Some studies have incorporated technologies to optimize clinical work processes and further enhance patient safety by utilizing automated clinical document classification [29].

In statistical keyword extraction strategies, statistical measures such as n-gram statistics, word frequency, and term frequency-inverse document frequency (TF-IDF) measures are used to detect keywords. Statistical keyword extraction strategies can be domain independent and do not require training data [30]. Xiao, Wang [31] proposed a text-classification-based model that utilized the Naive Bayes approach to solve patent document classification issues. The model uses NLP techniques to process the text, coupled with TF-IDF and JIEBA (a Chinese word segmentation tool) for keyword extraction and Naive Bayes as a classifier to obtain recall, F1-score, and accuracy of ≤93.9%. Jiang, Chen [32] proposed a text classification model to classify sensitive medical data utilized for medical information privacy protection. It uses diabetes text record samples sorted as sensitive and normal samples using JIEBA and TF-IDF for feature selection. In contrast, the k-nearest neighbor, support vector machine, and Naive Bayes were utilized as classification algorithms, achieving an accuracy of 92%.

Previous studies by Yeung [33] enumerated and described text classification steps: preprocessing, vector space model creation, dimensionality reduction (feature selection and projection), training of a classification function, and performance measurement. In their work, text classification, such as Boolean, TF-IDF, and entropy, was utilized to present multiple schemas of feature weighting. In addition, the authors explained three feature selection strategies: document frequency thresholding, information gain, χ²-statistic, and one feature projection method, and latent semantic indexing. Furthermore, they summarized and clarified six ML methods: Rocchio’s algorithm, Naive Bayes, k-nearest neighbor, decision tree, support vector machine, and ensemble learning, including bagging and boosting algorithms. Lastly, they described performance measures for binary, multiclass, and multilabel classification tasks.

In recent years, several studies have deployed different ensemble learning methods to predict mortality by using EHR data. For instance, Liu, Wu [34] used extreme gradient boosting to predict mortality in patients with acute kidney injury in an intensive care unit. Si, Du [35] proposed another method of crude clinical data and utilized deep learning to predict various clinical events, including in-hospital mortality, 24-h after admission. Ali et al. [36] proposed ML models for risk factor analysis and survival prediction in heart failure patients. The model identified six risk factors for heart failure patients. Another study [37] explored the effectiveness of ML classifiers in diagnosing and detecting cardiovascular diseases. The reviewed studies did not consider free text data in their analysis.

In a previous study [38], we developed an ensemble learning-driven framework to prevent misdiagnosis in patients with ACS. We compared the classification results of different boosting machine learning models (i.e., AdaBoost, gradient boosting machine, and extreme gradient boosting), and then compared their performance metrics. The results showed that our framework using boosting ensemble learning models outperformed the recall and F-measure metrics of other data processing levels (i.e., recall of 0.863 and F-measure of 0.863) while preserving a decent classification accuracy of 85.7%.

This study aimed to develop a reliable framework that can serve as a diagnostic support tool to prevent misdiagnoses. This study also advances our previous work by adding unstructured information to the previous models. The TF-IDF matrix features are extracted from free-text clinical narratives and added to the structured features, which are called merged data, throughout the entirety of this paper.

2. Methods

This section briefly introduces our preparation for the structured and clinical data. Our goal was to build and enhance an ensemble learning-driven framework, which will serve as a diagnostic support tool to prevent misdiagnosis among patients with clinical concerns for ACS. Many NLP tasks have benefited from the use of the TF-IDF. This concept is a statistical measure that assesses how essential a word is to a document in a collection of documents [39]. This is done by multiplying two metrics: how often a word appears in a document, and the inverse document frequency of the word across a set of documents. TF-IDF has been proven to benefit various NLP tasks [40]. We then briefly introduce the best-performing ensemble classifiers from a previous study, which were utilized to develop new models based on a new mixed dataset.
Next, we describe the architecture of the proposed enhanced model. The architecture of this study is illustrated in Fig. 1.

2.1. Data source and patient cohort

We used EHR data from a large urban emergency department (ED) located in Michigan, USA. The EHR data contained de-identified information of 362,138 patients who visited the ED between January 2017 and August 2020. The data records included clinical narratives along with quantitative and qualitative data, including demographic information, laboratory values, comorbidities, vital signs, and primary diagnoses coded using the International Classification of Diseases Clinical Modification, 10th Revision (ICD-10-CM) for patients with ACS.

Of the initial 362,138 patients who visited the ED, 16,096 with clinical concerns for ACS were selected for this study after necessary exclusions, of which 173 (1.1%), 47 (0.3%), and 15,876 (98.6%), respectively (see Fig. 2). The prevalence of ACS accounts for only approximately 1.8% of our data. The incidence of ACS is low because we considered only lower-risk patients without an initial instance of high troponin levels. The clinical diagnosis of ACS is less challenging in those patients who have high initial troponin values consistent with acute ischemia. Troponin testing is used to measure the levels of troponin T and I proteins in the blood. These proteins are released when the heart muscle is damaged (e.g., when cardiac ischemia occurs), and the greater the ischemic injury, the higher the troponin in the blood.

2.2. Data preprocessing

We identified a broad list of potential clinical features associated with ACS through a review of the medical literature and multiple discussions with medical professionals (e.g., emergency medicine physicians). From these lists, we selected 58 variables routinely recorded in the EHR, including demographic and clinical factors, to be included in the analysis. We also implemented NLP techniques to transform our clinical narrative into structured data fit for our research.

2.2.1. Unstructured data preprocessing: clinical narratives

Clinical narratives are descriptions of patients’ conditions and medical histories. In this research, clinical narratives were retrieved from the EHR discharge summaries using patient admission identification numbers. Discharge summaries are the primary method of communicating a patient’s care plan to the next provider [41]. Hence, they include rich information about the patient’s condition and treatment.

2.2.1.1. Feature engineering. This involves transforming the raw data into features that better represent the underlying structure of the data. The following steps were performed in this stage: (i) Text preprocessing: cleaning the data from distortion and useless information, such as punctuations, characters, and whitespaces. (ii) Transforming cases involves transforming all the upper and lower-case letters existing in the clinical narratives into lower-case letters, thereby eliminating homologous words that differ only in their case. (iii) Tokenizing: involves separating an entire sentence word by word for increased processing. (iv) Filtering stopwords: This algorithm is mainly utilized to eliminate unimportant words or words with no specific meaning, such as a, an, or the. This simplifies the computation using the following steps [42]:

2.2.1.2. Feature extraction. This involves transforming data into the desired form. We utilized the TF-IDF approach to extract relevant features from the data. TF-IDF is one of the most popular algorithms used in text-mining/processing research. Term frequency (TF) depicts the number of times a particular term appears in the text. Subsequently, inverse document frequency (IDF) was used to calculate the inverse probability of finding a word in a text [43].

\[ w_{ij} = tf_{ij} \times \log \frac{N}{df_{ij}} \]  

In equation (1), \( w_{ij} \) is the weight of word \( i \) in document \( j \), \( N \) is the number of documents in the entire corpus, \( tf_{ij} \) is the frequency of word \( i \) in document \( j \), and \( df_{ij} \) represents the number of documents contained in word \( i \) [44].

2.2.1.3. Feature selection. Feature Selection streamlines features to only those that contribute the most to the output [45]. SelectFromModel is a Random Forest feature selection model that selects features whose importance is greater than the mean importance [46]. Based on the features extracted after applying the TF-IDF, the SelectFromModel identifies the variables with the greatest contribution to enhancing the classification performance.

2.2.2. Structured data preprocessing: merged data

The structured EHR data include time of admission, demographics (e.g., age, sex, race, etc.), comorbidity information (e.g., cancer, chronic obstructive pulmonary disorder, etc.), primary diagnosis (e.g., NSTEMI, UA, etc.), and lab and chart values (e.g., cholesterol ratio, red blood cell distribution width, etc.). These features were selected on the basis of those used in similar studies.

2.2.2.1. Feature engineering. The following processes were performed during our analysis to achieve the intended objective: (i) One-hot encoding: substituting encoded categorical variables with newly generated binary variables to aid better classification. (ii) Removing redundant/duplicate data: noise, missing values, and discrepancies that adversely affect the quality of the dataset and alter the model’s predictive ability. Therefore, duplicate data rows and features with more than 70% of the missing values were excluded. (iii) Data imputation via Multiple Imputations using Chained Equations [47].

2.2.2.2. Feature selection. BorutaShap is applied when identifying the variables with the greatest contribution to enhancing classification performance [48]. BorutaShap is a wrapper feature selection technique that consolidates the Boruta [49] feature selection algorithm with SHAP.

![Fig. 1. Architecture of proposed methodology.](image-url)
Imbalanced classification issues occurred when the distribution of instances across classes was skewed (e.g., 173 cases with NSTEMI, 47 cases with UA, and 15,876 cases with non-ACS etiologies). These classifications are daunting for predictive modeling because ML algorithms are designed based on the assumption that the model has an equal chance of learning [51]. Because imbalanced training datasets violate this assumption, they lead to the development of models that exhibit poor performance, especially for the minority class (e.g., 173 cases with NSTEMI, 47 cases with UA, and 15,876 cases with non-ACS etiologies). The synthetic minority oversampling technique (SMOTE) was employed to address data imbalance, where instances from the minority class were randomly duplicated. SMOTE creates synthetic samples from the minority class using information available from the dataset [52]. A balance in the training dataset occurred because of the addition of these duplicated values to the minority class, thereby providing the model with an equal chance of learning. Applying SMOTE to an imbalanced dataset aid in the performance enhancement of ML algorithms in comparison with models without any balancing. This oversampling technique adjusted the ratios among the classes to achieve a ratio of 1:1:1 (i.e., 15,876:15,876:15,876).  

2.4. Prediction models  

We applied three learning algorithms to our analysis: logistic regression, AdaBoost, and linear discriminant analysis. Logistic regression (LR) is used as the baseline machine learning algorithm in this study [53]. AdaBoost (i.e., Adaptive Boosting) (ADA) tweaks subsequent weak learners to build a robust classifier. Although individual learners may be weak, the final model can converge into a strong learner [54]. The linear discriminant analysis (LDA) algorithm is a simple and effective method of classification and is intended for classification problems where the output variable is categorical. It also supports both binary and multiclass classification. LDA is also a feature extraction technique that computes transformations by maximizing the scatter between classes and minimizing that within classes [53]. The sample was divided into 70% training dataset and 30% testing dataset. The models were trained using the training datasets. Ten-fold cross-validation was used to validate the model.

2.5. Model evaluation  

The performance of each model was evaluated based on the area under the receiver operating characteristic (AUROC) curve and the F1 score. The F1 score is the harmonic of precision (positive predictive value) and recall (sensitivity), which complement the AUROC [56]. Because we are working with an imbalanced dataset in the study (i.e., 173 cases with NSTEMI, 47 cases with UA, and 15,876 cases with non-ACS etiologies), the F1 score provides a better explanation of the accuracy of our prediction models. The predictive accuracy of the model was determined using an AUROC curve. All analyses were conducted using Google Colab software (Python version 3.8).

3. Results

3.1. Unstructured data preprocessing  

There was a massive imbalance in our dataset, of which 173 (1.1%), 47 (0.3%), and 15,876 (98.6%) patients had NSTEMI, UA, and non-ACS, respectively. Fig. 3 depicts the length of the clinical notes by category to explore clinical narratives. Fig. 3 shows that, although the clinical note distribution length is different for each category, the difference is insignificant. We would have a problem if we had a wide variation in length between each category because the feature creation process may consider word counts. However, when creating the features using TF-IDF scoring, we normalized the features to avoid this. At this point, we could not conduct further investigation. Before the creation of any features from the raw text, it was necessary to clean the data and ensure that no misrepresentations were introduced into our proposed model. The categories in this context were the target variables (i.e., NSTEMI, UA, and non-ACS etiology). Given that ML models require numeric features and labels to provide predictions, creating a dictionary to map each label with a numerical ID is paramount. The mapping scheme is shown in Table 1.

To represent our text in the clinical notes, every row of the dataset will be a single corpus document. The TF-IDF vectors represent the column of the dataset (i.e., features) for this study. In the creation of features with this feature extraction process, we can choose the following parameters: (i) N-gram range: being able to consider unigrams, bigrams, trigrams, etc.; (ii) Maximum/Minimum Document Frequency (DF): in developing the word vocabulary, ignoring terms that have a DF strictly higher/lower than the given threshold is acceptable; and (iii) Maximum features: we can select the top features across the entire corpus. Table 2 lists the parameters selected for feature creation.

We expect bigrams and trigrams to enhance our model’s performance by considering words that tend to appear together in documents. We selected a Minimum DF value of 1.0 to eliminate infrequent words that do not appear in less than one document and a Maximum DF value...
of 1.0, which means that it ignores terms that appear in more than 100% of the documents. The selection of 300 as the maximum number of features was made to avoid the possibility of overfitting, which often emerges from many features in contrast to the number of training observations. After applying the SelectFromModel to the created features, 107 features were selected based on their contributions towards prediction, including abdominal pain, diabetes mellitus, shortness of breath, vital base pairs, alcohol use, nausea, and vomiting.

3.2. Structured data preprocessing

The patient (i.e., patients with NSTEMI, UA, or non-ACS etiologies) demographics and clinical outcomes are shown in Table 3, with a significance level of 5%. We found that the median age of all patients was 62.0 with first and third quartile of 53.0 and 71.0 years, respectively. Of the study patients, 45.1% were male and 85.1% were African Americans (or black). Comorbid conditions were common among patients, including abdominal pain, diabetes mellitus, shortness of breath, vital base pairs, alcohol use, nausea, and vomiting.

Fig. 3. Clinical notes length by category.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mapping schema for each class label.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category Name</td>
<td>Category Code</td>
</tr>
<tr>
<td>Non-ACS etiologies</td>
<td>0</td>
</tr>
<tr>
<td>NSTEMI</td>
<td>1</td>
</tr>
<tr>
<td>UA</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Selected parameters for feature creation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>Value</td>
</tr>
<tr>
<td>N-gram range</td>
<td>(2, 3)</td>
</tr>
<tr>
<td>Maximum DF</td>
<td>1.0</td>
</tr>
<tr>
<td>Minimum DF</td>
<td>1.0</td>
</tr>
<tr>
<td>Maximum features</td>
<td>300</td>
</tr>
</tbody>
</table>

of 1.0, which means that it ignores terms that appear in more than 100% of the documents. The selection of 300 as the maximum number of features was made to avoid the possibility of overfitting, which often emerges from many features in contrast to the number of training observations. After applying the SelectFromModel to the created features, 107 features were selected based on their contributions towards prediction, including abdominal pain, diabetes mellitus, shortness of breath, vital base pairs, alcohol use, nausea, and vomiting.

3.2. Structured data preprocessing

The patient (i.e., patients with NSTEMI, UA, or non-ACS etiologies) demographics and clinical outcomes are shown in Table 3, with a significance level of 5%. We found that the median age of all patients was 62.0 with first and third quartile of 53.0 and 71.0 years, respectively. Of the study patients, 45.1% were male and 85.1% were African Americans (or black). Comorbid conditions were common among patients, including abdominal pain, diabetes mellitus, shortness of breath, vital base pairs, alcohol use, nausea, and vomiting.

Fig. 3. Clinical notes length by category.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mapping schema for each class label.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category Name</td>
<td>Category Code</td>
</tr>
<tr>
<td>Non-ACS etiologies</td>
<td>0</td>
</tr>
<tr>
<td>NSTEMI</td>
<td>1</td>
</tr>
<tr>
<td>UA</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Selected parameters for feature creation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>Value</td>
</tr>
<tr>
<td>N-gram range</td>
<td>(2, 3)</td>
</tr>
<tr>
<td>Maximum DF</td>
<td>1.0</td>
</tr>
<tr>
<td>Minimum DF</td>
<td>1.0</td>
</tr>
<tr>
<td>Maximum features</td>
<td>300</td>
</tr>
</tbody>
</table>

Combining both data types (i.e., structured and unstructured data), the merged data consisted of 113 variables, with 107 variables derived from the clinical narrative word stem and the remaining six variables from the structured data. Furthermore, the merged data were split into independent training (70%) and testing (30%) sets to develop our classification framework. However, before undertaking our analysis, we must consider the data imbalance using SMOTE on the training set alone. When data distribution across known classes (e.g., NSTEMI:173 vs. UA:47 vs. non-ACS etiologies:15,876) is skewed, imbalanced classification issues arise.

Fig. 5 shows the data imbalance before and after implementing the SMOTE. Before data imbalance, NSTEMI, UA, and non-ACS etiologies were 173, 47, and 15,876, respectively, for the training set. After data balancing, the training sets for NSTEMI, UA, and non-ACS etiologies were 15,080, 15,080, and 15,080, respectively. To optimize the hyper-parameters for the different algorithms, we performed grid search cross-validation (GridSearchCV) to determine the optimal parameters for training the ADA, LDA, and LR algorithms (see Appendix 1). The combination of parameters was evaluated using the accuracy score as a performance metric and the best parameters with the highest accuracy scores were selected to develop the prediction framework.

Table 4 presents the results of the hold-out data (i.e., temporal validation) of all machine learning algorithms and a comparison between the modeling techniques across different data processing stages. As shown in Table 4, the first stage (i.e., modeling of raw data + feature engineering + feature selection) shows the performance of ADA, LDA, and LR on imbalanced data. LDA had the highest F1-score (0.980) compared with the other regression and ensemble classifiers. LDA had the highest accuracy (98.5%) and ADA had the highest AUC (1.000). In the second stage (i.e., modeling on raw data, feature engineering, feature selection, and oversampling techniques), we observed a decrease in the F1-score of ADA, LDA, and LR. It is also worth noting the significant improvement in the recall metrics of each classifier. The ADA model performed better at this stage with an F1-score of 0.943 and an accuracy of 94.3%. Accuracy is a better metric when working with a balanced dataset. However, most real-world problems are defined around an imbalanced dataset. While accuracy overlooks false positives and false negatives, the F1-score penalizes extreme recall or precision.

Comparing the accuracies reported by all models in Stage 1 (Table 4) with the baseline, only LDA outperformed the baseline classifier (LR).
Table 3
Baseline characteristics of patients.

<table>
<thead>
<tr>
<th>Variables</th>
<th>NSTEMI (N = 173)</th>
<th>UA (N = 47)</th>
<th>Non-ACS (N = 15,876)</th>
<th>Overall (N = 16,096)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (median [Q1, Q3])</td>
<td>63.0[57.0,72.0]</td>
<td>61.0[51.5,69.5]</td>
<td>62.0[53.0,71.0]</td>
<td>62.0[53.0,71.0]</td>
<td>0.071</td>
</tr>
<tr>
<td>Sex (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>53.8</td>
<td>51.1</td>
<td>54.9</td>
<td>54.9</td>
<td>0.834</td>
</tr>
<tr>
<td>Male</td>
<td>46.2</td>
<td>48.9</td>
<td>45.1</td>
<td>45.1</td>
<td></td>
</tr>
<tr>
<td>Ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do not know</td>
<td>4.0</td>
<td>–</td>
<td>2.9</td>
<td>2.9</td>
<td>0.484</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>2</td>
<td>1</td>
<td>2.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Not Hispanic/Latino</td>
<td>94.8</td>
<td>97.9</td>
<td>94.7</td>
<td>94.7</td>
<td></td>
</tr>
<tr>
<td>Race (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>80.3</td>
<td>61.7</td>
<td>85.2</td>
<td>85.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Other</td>
<td>4.6</td>
<td>–</td>
<td>4.4</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>15.0</td>
<td>38.3</td>
<td>10.4</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>Comorbidities (mean (SD))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>21 (16.8)</td>
<td>2 (5.1)</td>
<td>1396 (11.9)</td>
<td>1390 (11.9)</td>
<td>0.101</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>88 (70.4)</td>
<td>31 (82.1)</td>
<td>2240 (20.3)</td>
<td>2460 (21.1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>54 (43.2)</td>
<td>16 (41.0)</td>
<td>3236 (28.1)</td>
<td>3396 (28.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic heart disease</td>
<td>67 (53.6)</td>
<td>22 (56.4)</td>
<td>3147 (27.3)</td>
<td>3236 (27.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>91 (72.8)</td>
<td>29 (74.4)</td>
<td>5741 (49.4)</td>
<td>5861 (50.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>54 (43.2)</td>
<td>16 (41.0)</td>
<td>4288 (37.2)</td>
<td>4358 (37.3)</td>
<td>0.348</td>
</tr>
<tr>
<td>Hypertension</td>
<td>108 (86.4)</td>
<td>27 (69.2)</td>
<td>8677 (75.4)</td>
<td>8812 (75.5)</td>
<td>0.011</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>21 (16.8)</td>
<td>10 (25.6)</td>
<td>1022 (8.9)</td>
<td>1053 (9.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Obesity</td>
<td>38 (30.4)</td>
<td>5 (12.8)</td>
<td>2569 (22.3)</td>
<td>2612 (22.4)</td>
<td>0.035</td>
</tr>
<tr>
<td>Vital Signs and Lab values (mean (SD))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic blood pressure (mmHg)</td>
<td>148.8 (25.6)</td>
<td>140.4 (24.0)</td>
<td>139.2 (28.6)</td>
<td>139.3 (28.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diastolic blood pressure (mmHg)</td>
<td>84.4 (37.7)</td>
<td>80.1 (16.1)</td>
<td>80.4 (18.7)</td>
<td>80.4 (18.7)</td>
<td>0.017</td>
</tr>
<tr>
<td>Heart rate (bpm)</td>
<td>83.9 (19.7)</td>
<td>74.7 (13.4)</td>
<td>86.7 (20.7)</td>
<td>86.6 (20.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Temperature (°F)</td>
<td>98.1 (0.6)</td>
<td>98.2 (0.5)</td>
<td>98.1 (2.1)</td>
<td>98.1 (2.1)</td>
<td>0.965</td>
</tr>
<tr>
<td>Respiratory rate (breath/min)</td>
<td>21.1 (9.6)</td>
<td>18.3 (2.0)</td>
<td>19.6 (4.9)</td>
<td>19.7 (5.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Oxygen saturation (%)</td>
<td>97.3 (4.5)</td>
<td>98.0 (1.8)</td>
<td>97.2 (4.5)</td>
<td>97.2 (4.5)</td>
<td>0.490</td>
</tr>
<tr>
<td>Troponin (ng/L)</td>
<td>1.7 (4.4)</td>
<td>8.1 (1.1)</td>
<td>0.1 (1.2)</td>
<td>0.1 (1.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Brain natriuretic peptide (pg/ml)</td>
<td>447.4 (769.8)</td>
<td>121.1 (251.9)</td>
<td>402.8 (800.7)</td>
<td>402.4 (799.1)</td>
<td>0.081</td>
</tr>
<tr>
<td>Red blood cells (millions/μl)</td>
<td>4.4 (0.6)</td>
<td>4.3 (0.5)</td>
<td>4.3 (0.8)</td>
<td>4.3 (0.8)</td>
<td>0.046</td>
</tr>
<tr>
<td>White blood cells (thousands/μl)</td>
<td>9.3 (9.0)</td>
<td>7.1 (2.3)</td>
<td>8.2 (5.5)</td>
<td>8.2 (5.6)</td>
<td>0.023</td>
</tr>
<tr>
<td>Platelet count (thousands/μl)</td>
<td>226.9 (81.4)</td>
<td>226.9 (73.2)</td>
<td>230.0 (89.7)</td>
<td>230.0 (89.5)</td>
<td>0.876</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>38.9 (5.6)</td>
<td>38.5 (4.5)</td>
<td>37.3 (6.3)</td>
<td>37.4 (6.3)</td>
<td>0.005</td>
</tr>
<tr>
<td>Hemoglobin (g/dl)</td>
<td>12.8 (1.9)</td>
<td>12.8 (1.7)</td>
<td>12.2 (2.2)</td>
<td>12.3 (2.2)</td>
<td>0.002</td>
</tr>
<tr>
<td>Red cell distribution width (%)</td>
<td>15.2 (2.2)</td>
<td>14.6 (1.7)</td>
<td>15.6 (2.4)</td>
<td>15.6 (2.4)</td>
<td>0.001</td>
</tr>
</tbody>
</table>
The same notion applies to the second stage: ADA outperformed the baseline classifier. Nonetheless, ADA provides the best performance in Stage 2, and the primary objective of this study is not to promote any model, but to develop a decision support tool (i.e., the prediction framework) that is capable of handling data imbalance and feature engineering. Because ADA had the best performance in Stage 2, Fig. 6 shows the feature importance plot on the merged data, with troponin levels being the most important variable. The results from Stage 2 improved the performance of the models. Our results suggest that when data cleaning, feature engineering, and balancing techniques are implemented, the performance of the machine learning models is likely to increase.

Fig. 7 depicts the AUC curves for the trained models in Stage 2. In Fig. 7, we can observe that ADA has an AUC of 1.000 and 0.974, which means that it has a good separability measure among the three primary diagnosis classes. It is important to note that ADA had the highest AUC score in both data-processing stages. Next, we observed that the macro-average AUC curves for each model in Stage 2 were 0.98, 0.93, and 0.96, respectively. Therefore, the models performed well in classifying the positive class in the dataset after implementing the feature engineering, feature selection, and oversampling techniques. The optimal decision threshold value (i.e., the operational point on the AUC plot) was determined using Youden’s J index [57], which maximizes the difference between the true and false positive values.

4. Discussion

Cardiovascular disease accounts for approximately one-third of global deaths, with an estimated seven million deaths attributed to ischemic heart disease (IHD) [58, 59]. The World Health Organization states that ACS and sudden death are the leading causes of IHD-related fatalities, accounting for approximately 1.8 million deaths annually. According to the American Heart Association, an estimated 805,000 Americans experience annual ACS events. Of these cases, 605,000 are new attacks and 200,000 are recurrent attacks [60]. ACS treatment is especially vulnerable; globally, there have been reports of decreased ACS admissions accompanied by worsening in-hospital outcomes [61–63]. Healthcare avoidance and treatment delay are apparent [62, 63] and may further translate into poorer medium- or long-term cardiovascular outcomes, such as a higher incidence of heart failure. There is a need to predict the occurrence of ACS in patients to improve hospital resource planning and to develop a decision support system that can improve patient outcomes.

We developed a prediction framework that applies NLP to ED-assessed chief complaints in combination with structured data to detect early cases of ACS in patients with non-ACS etiology. We used three ML models, ADA, LDA, and LR, to classify NSTEMI, UA, and non-ACS. The best performance was achieved using ADA when the data were adequately engineered and preprocessed to address the class-imbalance issue. It is important to note that most ML models are not capable of handling class imbalances in each dataset. In our dataset, 98.6% of patients presenting to the ED had a non-ACS etiology. During prediction, ML models trained on imbalanced datasets often favor the majority class, leading to a better performance in the majority class than in the minority class. According to Ref. [64], developing models that are
insensitive to class distribution or utilizing oversampling techniques (e.g., SMOTE) will help address the class imbalance present in a given dataset. Our prediction framework (Stage 2), which combines feature engineering, feature selection, and SMOTE to address the imbalanced dataset, shows that the models can distinguish ACS patients (i.e., NSTEMI and UA) from non-ACS patients, as presented in Table 4. We used the F1-score to measure the performance of the models, as it is a better metric for imbalanced datasets. The results showed different performance scores for the algorithms.

Focusing on Stage 2 (balancing technique applied), we observed that the ADA model performed better than the other two models after feature engineering. In addition, Stage 2 shows a proper balance in the evaluation metrics compared with Stage 1 (i.e., when no balancing technique is applied) and a major improvement in the recall values of the models. One reason, in our opinion, is the ability of ADA classifiers to tweak multiple subsequent weak learners to build a robust classifier. AdaBoost uses what is called the decision stump. Decision stumps are decision trees that have one node and two leaves. AdaBoost uses multiple decision stumps, with each decision stump built on just one variable or feature [65]. This is unlike a random forest, in which decision trees use multiple variables to make a final classification decision. Although the ADA model was less prone to overfitting (i.e., a decrease in prediction performance), we were careful not to overestimate the complexity of the model during tuning. ADA models have provided varying performance results in different applications including bioinformatics [66], hotspot detection [67], medical informatics [68], and other domains such as image analysis [69]. To the best of our knowledge, this is the first study to use a machine learning approach to classify patients with ACS from those with non-ACS etiology using structured and unstructured data.

In our study, several factors were identified as independent predictors to distinguish patients with ACS from non-ACS patients. Among the structured data, troponin level, glucose level, a history of myocardial infarction, and systolic blood pressure emerged as pivotal indicators. Notably, from clinical narratives, which are inherently unstructured data sources, documentation of distal pulses, a number of additional variables significantly improved predictive modeling. These include clinicians bringing attention to findings such as visual disturbance, the character of chest pain, the patient’s blood pressure, oxygen saturation, and body mass index. While some of these factors may also be found as structured variables, the clinician emphasis in unstructured notes was additive to the models. Although previous studies have leveraged patient and electrocardiogram data for ACS prediction [70-74] to our knowledge, no research to date has integrated clinical narratives as a data source for this purpose. Our innovative approach of blending structured data with the depth and nuance of clinical narratives fills this gap in the literature and provides a comprehensive framework for ACS prediction, underscoring the rich insights gleaned from diverse data streams.

Applied in real time, our ML framework can help identify patients with ACS at the time of presentation. These machine learning models can be expanded to other outcomes of interest such as personalized risk prediction models and other clinical outcomes. A study by Jokardarabi et al. [75] used AdaBoost, among other machine learning models, to predict the in-hospital mortality of acute myocardial infarction with an AUC of 0.987. Machine learning has been utilized for the early prediction of in-hospital cardiac arrest in patients with acute coronary syndrome [76]. They reported an accuracy of 96% and an F1-Score of 80% [76]. The advantages of our study include the focus on chest pain patients, development of a predictive framework, and evaluation of the framework’s performance using accuracy, recall, precision, F1-score, etc.
and AUC to ensure proper reporting of the analysis performed on imbalanced datasets.

Our study has several limitations. First, some clinical variables had missing values owing to administrative errors. Missing data points were imputed using multivariate imputation with the chained equations approach, introducing bias if the missingness was not random. Second, our model was limited to features readily available in a hospital’s EHR system. Features not captured or retrieved from the EHR, such as patients being transferred from another hospital, were not included in the study. Finally, using ICD-10 CM codes for cohort identification may introduce noise. Nevertheless, this noise did not affect the findings of the present study. With slight modifications to our model, other EDs could use the prediction framework to identify significant predictors in their data.

5. Conclusion

In this study, we combined unstructured and structured data to build a framework that will serve as a diagnostic support tool for preventing misdiagnoses among patients with clinical concerns for ACS. We extracted the TF-IDF features from each patient admission discharge summary for unstructured data. Our findings suggest that adding features from unstructured data to a richer data dimension enhances the performance of the proposed framework. In the future, we intend to design deep neural network ensembles that have the potential to further improve model performance and investigate unstructured sequential data using other cutting-edge models such as recurrent neural networks and long short-term memory (LSTM) methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.healthc.2023.100249.

References

[38] A. Hakim, et al., Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF)


[48] E. Keany, Is This the Best Feature Selection Algorithm It’s the way I wish someone explained it to me-ab81cc69ef30. (Accessed 2 September 2023).


[68] W. Abdelkader, et al., A deep learning approach to refine the identification of high-quality clinical research articles from the biomedical literature: protocol for algorithm development and validation, JMIR research protocols 10 (11) (2021), e20936.


