



Published in final edited form as:

Birth Defects Res. 2025 March ; 117(3): e2451. doi:10.1002/bdr2.2451.

Machine Learning and Natural Language Processing to Improve Classification of Atrial Septal Defects in Electronic Health Records

Yuting Guo¹, Haoming Shi², Wendy M. Book^{3,4}, Lindsey Carrie Ivey⁴, Fred H. Rodriguez III³, Reza Sameni¹, Cheryl Raskind-Hood⁴, Chad Robichaux¹, Karrie F. Downing⁵, Abeed Sarker¹

¹Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, Georgia, USA

²Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

³Department of Cardiology, School of Medicine, Emory University, Atlanta, Georgia, USA

⁴Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

⁵National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Abstract

Background: International Classification of Disease (ICD) codes can accurately identify patients with certain congenital heart defects (CHDs). In ICD-defined CHD data sets, the code for secundum atrial septal defect (ASD) is the most common, but it has a low positive predictive value for CHD, potentially resulting in the drawing of erroneous conclusions from such data sets. Methods with reduced false positive rates for CHD among individuals captured with the ASD ICD code are needed for public health surveillance.

Methods: We propose a two-level classification system, which includes a CHD and an ASD classification model, to categorize cases with an ASD ICD code into three groups: ASD, other CHD, or no CHD (including patent foramen ovale). In the proposed approach, a machine learning

Correspondence: Yuting Guo (yguo262@emory.edu).

Author Contributions

Yuting Guo: conceptualization, methodology, investigation, data curation, writing – original draft, review and editing, analysis.

Haoming Shi: conceptualization, methodology, investigation, data curation, writing original draft, review and editing, analysis.

Wendy M. Book: conceptualization, methodology, investigation, data curation, writing – original draft and review and editing, preparation, supervision, project administration, funding acquisition. **Lindsey Carrie Ivey:** conceptualization, methodology, investigation, resources, data curation, writing original draft and review and editing, analysis, supervision, project administration.

Fred H. Rodriguez III: conceptualization, methodology, investigation, data curation, review and editing, preparation, supervision.

Reza Sameni: conceptualization, methodology, investigation, data curation, writing – original draft, review and editing, analysis.

Cheryl Raskind-Hood: review and editing. **Chad Robichaux:** data curation. **Karrie F. Downing:** conceptualization, review and editing. **Abeed Sarker:** conceptualization, methodology, investigation, data curation, writing – original draft, review and editing, analysis, supervision.

Conflicts of Interest

The authors declare no conflicts of interest.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

model that leverages structured data is combined with a text classification system. We compare performances for three text classification strategies: support vector machines (SVMs) using text-based features, a robustly optimized Transformer-based model (RoBERTa), and a scalable tree boosting system using non-text-based features (XGBoost).

Results: Using SVM for both CHD and ASD resulted in the best performance for the ASD and no CHD group, achieving F_1 scores of 0.53 (± 0.05) and 0.78 (± 0.02), respectively. XGBoost for CHD and SVM for ASD classification performed best for the other CHD group (F_1 score: 0.39 [± 0.03]).

Conclusions: This study demonstrates that it is feasible to use patients' clinical notes and machine learning to perform more fine-grained classification compared to ICD codes, particularly with higher PPV for CHD. The proposed approach can improve CHD surveillance.

Keywords

atrial septal defect; birth defects; congenital heart disease; natural language processing

1 | Introduction

Electronic health records (eHRs) are promising resources for population-based surveillance studies and secondary analysis of existing data. *International Classification of Diseases, Clinical Modification* 9 and 10 (ICD-9-CM and ICD-10-CM) and *Current Procedural Terminology* (CPT) codes are used to classify diagnoses, procedures, and outcomes of interest, with high variability in accuracy of these codes for identification of the problem of interest. Many studies have explored machine learning (ML) algorithms, including supervised learning methods like support vector machines (SVMs), decision trees, random forests, and deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), particularly for disease classification tasks in clinical settings (Miotto et al. 2016; Rajkomar et al. 2019). These approaches leverage clinical notes, eHRs, lab results, and imaging data to predict diagnoses or identify patient phenotypes. In particular, ML-driven classification of diseases into ICD codes has garnered substantial attention. Several studies have focused on automating ICD code assignment from clinical notes, leveraging techniques such as natural language processing (NLP) combined with ML to capture the semantics of complex medical language. For instance, Masud et al. (2023) used CNNs on clinical text to automatically classify diagnoses to corresponding ICD codes, showing promising results in accuracy improvement over traditional rule-based approaches. Similarly, Xie and Xing (2018) proposed a model using both CNNs and RNNs to handle the hierarchical nature of ICD codes and capture contextual dependencies, achieving state-of-the-art performance in ICD classification tasks. Recent systematic reviews have highlighted the growing use of NLP for automated ICD coding of discharge summaries (Kaur et al. 2023; Stanfill et al. 2010; Shi et al. 2023). These reviews identified various data sets, NLP techniques, and evaluation metrics used in the field. These findings suggest that NLP-based approaches offer significant potential for improving the accuracy and efficiency of automated ICD coding systems.

Classification of congenital heart defects (CHDs), based on CHD-related ICD-9-CM and ICD-10-CM code groups 745.xx–747.xx and Q20.x–Q26.x (Glidewell et al. 2018), respectively, have variable success in identifying CHD (Khan et al. 2018; Broberg et al. 2015) primarily because the most common code within these code groups, making up an estimated of 25% cases in data sets (Xie and Xing 2018), is 745.5/Q21.1, which codes for *both* secundum atrial septal defect (ASD), a true heart defect, *and* patent foramen ovale (PFO), a common and normal variant in the general population (Rodriguez et al. 2018, 2022). ICD code 745.5/Q21.1 has a positive predictive value (PPV) as low as 30% for having *any* CHD (Ivey et al. 2023), and among those that do have CHD, a small percentage have a different CHD other than ASD (Rodriguez et al. 2018). Some ICD codes in the CHD code group are accurate, and the PPV of other CHD codes can be improved with ML techniques (Shi et al. 2023), yet the ICD code 745.5/Q 21.1 remains problematic given its dual function to code for both a heart defect and a normal variant, leading to a high number of false positives (FPs) in ICD code defined CHD data sets. As the most common code in a CHD data set, it is important to identify techniques to further improve the accuracy of distinguishing true CHD from FP normal variants within an ICD code defined CHD data set.

Consequently, erroneous conclusions about the CHD population as a whole may be drawn from studies inclusive of these codes. Conversely, exclusion of isolated 745.5/Q21.1 from studies of individuals with CHD limits understanding of healthcare utilization and outcomes unique to those with secundum ASD. CHD patients with ASD may experience complications related to the ASD including supraventricular tachyarrhythmias (Kauling et al. 2024), pulmonary arterial hypertension (Deniwar et al. 2023), and heart failure. Given the need for public health surveillance of CHD across the lifespan in general, more accurate classification of cases identified with ICD 745.5/Q21.1 is needed for a more inclusive and accurate data set to reach valid conclusions about the population as a whole.

In 2023, ICD-10-CM created a separate code, Q21.12, for PFO; however, while the transition to this new code is taking place over time, the utility of existing, longitudinal data available from birth defects surveillance currently relies on our ability to distinguish true CHD from normal variants within the data. Methods developed for automating this task may also be applied to other similar classification tasks (aside from ASD/PFO) where ICD codes are not accurate in identifying the population of interest.

ML and NLP techniques may improve the accuracy of identifying CHD in a specific cohort based on data from eHRs. Shi et al. (2023) developed a ML model that improved the performance of CHD identification using the structured features (variables) extracted from eHRs. Clinical notes from eHRs also contain rich, unstructured information that often provides essential context beyond structured data, such as lab results or ICD codes alone. NLP techniques enable the extraction of meaningful insights from free-text descriptions, capturing nuances in symptoms, diagnoses, and treatment details. This process improves classification accuracy, particularly in complex cases where multiple comorbidities or unclear symptomatology make diagnosis challenging. Studies have shown that NLP-based models can enhance the identification of diseases such as diabetes and cardiovascular conditions by effectively processing clinical notes and other unstructured data sources, providing contextually relevant information that improves model predictions (Banerjee

et al. 2023; Khalifa and Meystre 2015). Using eHR notes, it has also been previously demonstrated that NLP can detect patients who have had a Fontan operation for rare single ventricle heart defects better than ICD codes can (Guo et al. 2023). ASD presents a unique challenge in that text notes may variably include both ASD and the FP PFO for the same case.

With the goal of more accurately identifying *CHD cases* among individuals *identified by code 745.5/Q21.1*, by *detecting ASD among those with this code* and to improve the overall accuracy of an ICD code defined CHD data set that has previously been improved through ML techniques, the current study sought to create a classification system integrating ML and NLP techniques to identify (1) CHD and (2) ASD. The proposed approach can be replicated for similar problems within the broader medical space where variability in clinical concepts exists within ICD or similar codes.

2 | Materials and Methods

2.1 | Study Population and Data Collection

Individuals aged 0–55 years with at least one CHD ICD code and 2010–2019 encounter level data were identified as part of a parent CDC-funded CHD surveillance project (DD19-1902A and 1902B). Two validated cohorts were used for this study—an administrative data cohort from a large statewide adult tertiary referral healthcare system and a distinct statewide pediatric tertiary referral healthcare system, hereafter referred to as AHS/PHS (Ivey et al. 2023), and a Society for Thoracic Surgeons (STS) database. The adult and pediatric healthcare systems have distinct eHRs, charting styles, and specialists. A complete description of the AHS/PHS cohort has been previously described (Ivey et al. 2023). Briefly, 1497 individuals born between January 1, 1955 and December 31, 2019 who had CHD-related healthcare encounters (i.e., encounters with at least one of 87 ICD-9-CM or ICD-10-CM codes in Table S1) in either the AHS or PHS between January 1, 2010 and December 31, 2019 were identified and available clinical *notes* were abstracted to validate and classify their CHD status as one of three mutually exclusive groups: any ASD, other CHD only, no CHD (including no CHD but with PFO) (Ivey et al. 2023). Additional details on the validation data set, the process of case validation, and interobserver and intraobserver reliability (exceeded 95%) for the validation data were previously reported in Ivey et al. (2023). The STS research database includes data on pediatric patients who received cardiothoracic surgeries at the PHS (some patients were identified in both data sets); 9811 patients with CHD-related surgical encounters between January 1, 2010 and December 31, 2019 were identified with validated CHD status. The STS database captures children undergoing cardiac surgery, 95% of whom have CHD, with data on the specific diagnosis entered by clinicians on the cardiothoracic surgical team. Encounter-level raw data for all cases were available from the parent CHD surveillance project for 2010–2019. Cases were similarly classified into one of the three groups: ASD, other CHD, or no CHD (includes PFO without CHD).

For this analysis, only cases with code 745.5 or Q21.1 and text notes were included. There were 3778 cases in the data set who met the inclusion criteria (3181 from STS, 522 from AHS/PHS, and 75 from both). Finally, the cases in the data set were categorized into one

of the following three groups, based on cardiology provider-entered diagnoses in STS or diagnoses validated through medical record review by medical experts: individuals with an ASD ($N = 1395$), individuals with other CHD ($N = 2027$), and individuals with no CHD ($N = 356$). For each individual, clinical notes from all encounters were compiled and used as input to the model. Individuals with ASD, other CHDs, and no CHD had an average of 65, 72, and 155 clinical notes, respectively, with an average of 1517, 1533, and 798 words per note.

2.1.1 | Race and Ethnicity—Race and ethnicity were based on data recorded in the eHR and reported to assess generalizability of the findings to other data sets and potential biases in the system. Individuals could have more than one race recorded. Race was categorized as White, Black, or other race (due to small numbers, Asian, American Indian/Native American, Native Hawaiian/Pacific Islander, and multiracial were analyzed together [excluding Black multiracial who were instead categorized as Black]). Racial groups of small size ($n < 20$ in the AHS/PHS) were combined into the “Other” race category. Ethnicity was classified as Hispanic or non-Hispanic.

2.1.2 | Statistical Analysis for Cohort Characteristics—First encounter was defined as the first available encounter during the 2010–2019 surveillance period. Wilcoxon rank sum test was used to compare mean age at first encounter between the AHS/PHS cohort and the STS cohort. Sex, race, ethnicity, and other characteristics were defined as in Ivey et al. (2023) and compared using Pearson’s chi-square test. Proportion with a CHD was calculated as cases with a validated CHD (ASD or other) divided by all cases with a 745.5/Q21.1 ICD code. PPV of ICD codes 745.5/Q21.1 for an ASD was calculated as cases with a validated ASD divided by all cases with only ICD codes 745.5/Q21.1. Proportion with a CHD and PPV were also compared between the AHS/PHS and STS data sets using Pearson’s chi-square test.

2.2 | Classification System Development and Evaluation

Given the goal of improving accurate identification of (1) CHD cases in ICD-code based CHD data sets, and (2) identifying ASD within those data sets, we approached the problem with a two-level system. This approach allows us to analyze each component separately, rather than training a single three-way model to classify all three groups (any ASD, other CHD only, and no CHD).¹ Figure 1 shows the system framework for training CHD and ASD classification models and testing (evaluating) model performances. To reduce the risk of overfitting and provide a more robust evaluation of the model performance compared to a single train-test split, fivefold cross-validation was applied to evaluate the model performance, as shown in Figure S1. Specifically, the data set was randomly divided into five equal parts. For each CHD and ASD classification, the model was trained and tested five times during cross-validation, each time using a different fold for testing and the remaining four folds for training.² During training, a CHD classification model and an ASD classification model were independently trained on the training set. During testing, the

¹Although the two-level system may carry a higher risk of error propagation, our preliminary experiments showed no significant difference in performance between the two-level approach and the three-way classification model.

procedure included two classification steps. The test cases were first predicted by a CHD classification model. The cases predicted as having CHD during CHD classification were then processed through an ASD classification model to see if they were ASD or other CHD. Cases not identified as CHD during CHD classification were categorized as no CHD and no further action was taken.

All models were trained on AHS/PHS and STS cohorts combined. For the main analysis, testing was performed only on the AHS/PHS, because the AHS/PHS data set has a lower baseline PPV that would necessitate ML improvements (PPV 30.1%), unlike the STS data (PPV 94.7%). For comparison, we provide performance results when using both cohorts for testing to compare how the models perform on those cases combined relative to the performance on AHS/PHS cases alone.

2.3 | Models for CHD and ASD Classification

For both CHD and ASD classification, we compared the performance of two NLP models: SVM (Chang and Lin 2011) using text-based features (Khalifa and Meystre 2015) and a robustly optimized Transformer-based model (RoBERTa) (Liu et al. 2019). For CHD classification, we also compared the performance of a previously trained ML model, namely a scalable tree boosting system (XGBoost) (Chen and Guestrin 2016) that incorporated only non-text features (i.e., features derived from structured data and not clinical notes). As a supplemental analysis, we compared how all three models perform in CHD classification alone (i.e., stopping after the first step in the testing framework of Figure 1).

2.3.1 | SVM Model for CHD and ASD Classification—SVMs are often used when dealing with large feature spaces, making them a popular choice for text classification. To represent the clinical text notes as features, Term Frequency-Inverse Document Frequency (TF-IDF) for n grams (sequences of n words) was used, and in the current study, 1, 2, 3, and 4 grams were employed. TF-IDF is a statistic that measures the importance of a word in a document or set of documents. TF represents how often a word appears in a document (such as a clinical text note), while IDF represents how rare the word is across all documents in a training data set. TF-IDF helps to determine the significance of words in a document compared to the entire set of documents. Consequently, TF-IDF vectors assign higher numerical values to unique n grams in each document and lower values to those that appear uniformly across all documents. Stopwords were excluded, and only the top 1000 features, ranked by term frequency, were used for computing TF-IDF. In the training process, grid search was employed with the hyperparameters shown in Table S2. Furthermore, a class weighting strategy was implemented that automatically adjusts weights based on the inverse proportion of class frequencies in the input, so that the majority class receives a lower weight compared to the minority class during training to address imbalances in the data set. The equation for computing the class weight was described in Equation (S1).

²We implemented a nested cross-validation approach. During training, we performed fivefold cross-validation on the training set. In each fold, the data was split into a sub-training set and a subtest set. We then used grid search to identify the optimal hyperparameters by evaluating performance across the subtest sets.

2.3.2 | Transformer-Based Model for CHD and ASD Classification—

Transformer-based models like RoBERTa have shown outstanding performance in many NLP tasks. Unlike traditional methods that extract features or generate n grams, RoBERTa splits clinical text notes into word pieces or tokens. Each token is then encoded into a vector, and these vectors are combined to form a vector representation of the clinical narrative. Unlike n -gram vectors that are sparse, the vectors generated by transformer-based models are dense and capture the context of each token. We chose RoBERTa from the many transformer-based models currently available since it has been demonstrated to be the best in past benchmarking studies, where RoBERTa achieved the top or comparable performance compared to the transformer-based models trained on clinical text (Liu et al. 2019; Guo et al. 2020, 2022). However, RoBERTa has a limitation of only being able to handle texts up to 512 tokens in vector format, and many clinical text notes exceed this length. To overcome this limitation, a sliding window strategy was employed for the current data, where long clinical text notes were split into multiple partial clinical text notes. Each partial clinical text note is represented as an individual document with a sliding window of 512 tokens. The model was applied independently to each subdocument, allowing for independent classifications. After obtaining classifications for each subdocument, a final classification approach was determined by taking the majority vote over all the subsequences. No cleaning was performed on the text. The average number of subdocuments for individuals with CHD, no CHD, ASD, and no ASD were 207, 241, 193, and 221, respectively. For specific RoBERTa, hyperparameters and technical details applied in the current study, see Table S3.

2.3.3 | XGBoost for CHD Classification—In our previous work, XGBoost outperformed other tested models (logistic regression, Gaussian Naïve Bayes [NB], and Random Forest) for similar CHD classification efforts (Shi et al. 2023). Therefore, for CHD classification in this project, an XGBoost model was trained using non-text features from the broader AHS/PHS data set of 1497 validated cases with CHD codes, and a legacy clinical cardiac database of 1213 validated cases. A variety of features (predictive variables), totaling 33,460, were initially created for each patient by summarizing and categorizing their ICD-9-CM codes, ICD-10-CM codes, and CPT codes across all medical visits in the surveillance period. Using the same training process described by Shi et al. (2023), feature selection was applied to identify and retain the subset of the most relevant features ($n=2105$) that best identified a true CHD, and XGBoost models with these selected features were trained on the AHS/PHS and legacy clinical cardiac data. This model was compared to NLP models that used text features for CHD classification and was evaluated on the same validation data as the NLP models (i.e., the test set of the AHS/PHS cohort with ASD/PFO ICD codes).

2.4 | Post-Classification Analysis

2.4.1 | Model Performance—Model performances were measured by the precision (PPV), recall (sensitivity), and F_1 score (harmonic mean of precision and recall) metrics for each of the three groups (ASD, other CHD, no CHD). The F_1 score served as the primary metric for comparison to ensure that neither precision nor recall is optimized at the expense of the other. Bootstrap resampling was used to compute 95% confidence intervals for the F_1 scores (Efron 1979). Equations for computing these metrics are shown in Equation (1). The

mean and variance of performances across the five folds were computed. As supplemental analyses, we further assessed model performance in identifying individuals with any CHD (ASD and other CHD combined).

$$\text{Recall (sensitivity)} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (1)$$

$$\text{Precision (PPV)} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$F_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.4.2 | Receiver Operating Curve (ROC) and Precision Recall (PR) Curve—

For CHD classification, we used two tools to assess the impact of having an imbalanced distribution of cases with and without CHD on model performance: the (ROC) and the PR curve. The ROC curve portrays the model's ability to distinguish between TP and FP by plotting the TP rate versus the FP rate at various thresholds. Additionally, a NB classifier using the same features as SVM was implemented as a baseline model, and a “no skill” classifier that performed random binary classification was plotted as a diagonal line for comparing models' capabilities. The PR curve visualizes the precision score versus the recall score at different threshold levels. Both graphical representations provide an assessment of the model's overall discriminatory capability. Area under curve (AUC) scores for the ROC and PR curves were also computed. ROC and PR curve analyses were performed for SVM and XGBoost for CHD classification. RoBERTa was not included because RoBERTa did not produce probability vectors, so RoBERTa's performance could not be evaluated at distinct thresholds.

2.4.3 | Error Analysis—For both CHD and ASD classification, confusion matrices were generated for all three models (SVM, XGBoost, and RoBERTa) to uncover patterns in misclassifications. Confusion matrices provide systematic and structured breakdown of model predictions and categorize them into four distinct classes: true positives (TPs), true negatives (TNs), FPs, and false negatives (FNs). This delineation allows for a comprehensive understanding of model error patterns and, in turn, facilitates the identification of inherent weaknesses. This tool assesses how well models perform overall and allows the study of the ability of the model to identify true cases (TP and TN) and its tendency to make accurate classification/identification errors (FP and FN). This analysis provides insights into model performance and identifies areas for further model refinement or optimization.

3 | Results

3.1 | Cohort Characteristics

Characteristics of the analytic cohorts are noted in Table 1. The AHS/PHS cohort was older ($p < 0.001$), had a lower proportion with true CHD ($p < 0.001$), and had lower PPV for ICD codes 745.5/Q21.1 ($p < 0.001$) in comparison to the STS cohort. There were more Hispanic patients in the STS cohort compared to the AHS/PHS cohort ($p < 0.001$), whereas the AHS/PHS cohort had more individuals with unknown race ($p = 0.03$) and unknown ethnicity ($p < 0.001$) compared to the STS cohort. In the AHS/PHS cohort, overlapping cases were identified, thus they are reported together in Table 1. To evaluate potential differences in this combined AHS/PHS data set, all overlapping cases were analyzed as AHS. Mean age at first encounter was higher in the AHS compared to PHS (43.6 vs. 3.0 years, $p < 0.001$). Female gender was more common in the AHS compared to PHS (61% vs. 45%, $p < 0.001$), whereas Hispanic ethnicity was more common in PHS (10.0% vs. 4.8%, $p < 0.001$), and there was no significant difference in race. The PPV for code 745.5/Q21.1 was 38.3% for PHS compared to 25.4% for AHS ($p < 0.05$).

3.2 | Classification Results

Table 2 presents the evaluation metrics, including precision, recall, and F_1 scores, for the three groups (ASD, other CHD, no CHD) separately for the AHS/PHS cohort. (results for the AHS/PHS + STS cohort are in Table S4). For the ASD group, the system using SVM for CHD classification and RoBERTa for ASD classification resulted in the highest overall precision score (0.90) but at the expense of recall (0.18); in other words, 82% of true ASD cases are excluded from this group using this combination. The system using SVM for both CHD and ASD classification achieved the highest F_1 score (0.53), demonstrating a better balance between precision (0.83) and recall (0.39). For the other CHD group, the SVM-based system for CHD and ASD classification had the best precision (such that 33% actually had CHD other than ASD), but the XGBoost-SVM system achieved the highest F_1 score (0.39). For the no CHD group, the system using SVM for CHD classification achieved comparable precision (0.69) but significantly higher recall (0.92) than other models, contributing to a higher F_1 score (0.78). When comparing model performance for CHD classification alone (Table S5), XGBoost had the highest F_1 score (0.69) in the AHS/PHS data but the SVM system had the highest precision; of cases classified as having CHD with SVM, 84% actually had a CHD (i.e., a recall of 0.84), compared to 45.7% when relying on ICD-codes alone (as noted in Table 1).

3.3 | Post-Classification Analysis

3.3.1 | ROC and PR Curve Analysis—Figure 2 illustrates the ROC and PR curves for SVM, XGBoost, NB, and “no skill” models for CHD classification, respectively. For both curves, SVM and XGBoost achieved substantially higher AUC scores compared to the NB classifiers, demonstrating that these models effectively identify cases with CHD. In both curves, the AUC score for the SVM model is slightly higher than that for the XGBoost model (consistent with results in Table 2). These findings suggest that the SVM and XGBoost models were not substantially influenced by the data imbalance.

3.3.2 | Error Analysis—For CHD classification (Figure 3a), the SVM model exhibited an increase in TNs and a decrease in FPs when compared with results from the RoBERTa and XGBoost models. The difference between TNs and FPs for SVM was significantly larger than that for RoBERTa and XGBoost. However, the difference between TPs and FNs for SVM was significantly smaller than that for XGBoost and comparable to RoBERTa. For ASD classification among the entire AHS/PHS cohort (Figure 3b), both models exhibited substantially more TNs than TPs. The difference between TNs and TPs for SVM was smaller than that for RoBERTa, but the difference between TPs and FNs for SVM was significantly smaller than that for RoBERTa. Overall, SVM outperformed RoBERTa for ASD classification for individuals with CHD.

4 | Discussion

Our CHD/ASD classification approach, using complementary methods of ML to improve the accuracy of the overall cohort within the CHD ICD code group followed by NLP to further improve accuracy of the problematic subgroup 745.5/Q 21.1 within the CHD code group, addresses the problem of detecting patient cohort subsets that may not be covered by unique ICD codes, thus improving the accuracy of the larger cohort to accurately reflect the population of interest. Results from our proposed two-level model, which leverages NLP and a ML model trained on structured data, demonstrate that it is a promising scalable solution for identifying and more accurately distinguishing individuals with CHD and ASD, individuals with CHD other than ASD, and individuals without CHD compared to ICD codes alone. This methodology can improve both inclusivity and accuracy of CHD surveillance data sets. Despite the high prevalence of ASD/PFO codes in administrative data (Rodriguez et al. 2018), many analyses using clinical and administrative data opt to exclude individuals with only these codes due to low PPV, and likely exclude 100% of true ASD cases (who have no other CHD) (Burchill et al. 2018; Lui et al. 2022; Gurvitz et al. 2020; Downing et al. 2022; Agarwal et al. 2019). As an alternative to this exclusion, among people with these codes and clinical notes available, our classification system improved the identification of true ASD cases from 30% to 83%, while retaining nearly 40% of all true ASD cases in the data set. The use of the F_1 score as the primary metric for system comparison ensured that equal weight was given to precision and recall—either can be optimized with significant cost to the other, resulting in inflated estimates of performance (e.g., a precision-optimized system may achieve 100% precision but only detect a small proportion of relevant patients). In application settings, investigators can assign weights based on the priority of the metric by using the F_β score, where $\beta > 1$ favors recall and $\beta < 1$ favors precision.

The results have also revealed several key insights into the performance of different classification models, namely SVM, RoBERTa, and XGBoost, in identifying true CHD and true ASD among individuals with ASD/PFO diagnosis codes. Different combinations of models had different strengths in precision and recall, depending on the classification group (ASD, other CHD, or no CHD) and the data set. Overall, the system using SVM for both CHD and ASD classification consistently achieved the highest F_1 scores for the ASD and no CHD groups, and the XGBoost-SVM system achieved the highest F_1 score for the other

CHD group. In the AHS/PHS data set, those identified by the SVM-SVM system as having an ASD account for 39% of all true ASD cases in the data, and of those classified as having ASD by the system, 83% actually have a true ASD. This is a substantial improvement compared to relying on the ICD-9-CM and ICD-10-CM code for ASD, of whom only 30.1% actually had a true ASD in the structured data. More accurate classification of ASD and PFO in eHRs would improve our understanding of these two distinct conditions which share an ICD code. Although a new ICD code for PFO, Q21.12, recently went into effect on October 1, 2023, historical databases and administrative data sets will not contain this distinction between ASD and PFO.

Through error analysis, the divergence in performance for CHD classification could be linked to the distinctive use of n -gram features by SVM. The incorporation of n -gram features allowed SVM to potentially capture essential words or phrases within the clinical notes, enhancing its ability to identify cases without CHD. This suggests that the textual content within clinical notes may contain valuable linguistic markers that were effectively harnessed by the n -gram feature representation. In contrast, RoBERTa leveraged document embedding, while the XGBoost model relied on meta features. In comparing these approaches, it appeared that the straightforward inclusion of plain text features from clinical notes by the SVM may offer a more potent means of representing the cases without CHD. This underscores the importance of considering feature extraction methods in the context of the specific characteristics of the data under examination. The confusion matrix further revealed that SVM was better at striking a balance between TPs and TNs, while RoBERTa was better at identifying the negative class than the positive class for ASD classification. This suggests that the specific indicators for ASD might not be distributed throughout clinical text notes but could be concentrated in specific words or phrases. It is also possible that RoBERTa, a language model pretrained on generic web content such as English Wikipedia and user-generated comments from Reddit (Liu et al. 2019), might not capture the linguistic nuances associated with less common cardiovascular conditions. Pretraining models with problem-specific data may lead to better performances, as observed in prior works (Guo et al. 2023; Downing et al. 2022; Agarwal et al. 2019).

Overall, results from the current study demonstrated the effectiveness of a rigorously constructed and validated ML and NLP pipeline for accurately classifying rare patient cohort subsets in the context of CHDs. This methodology has the potential to create more inclusive and more accurate CHD data sets from eHRs, and may also be adapted for solving other similar problems within the broader medical domain. The key strengths of the current study can be summarized as follows:

- This study takes an innovative approach to improve CHD and ASD classification by leveraging clinical text notes with carefully crafted annotation guidelines, contributing to a high-quality, labeled data set for these conditions. While prior studies have also used clinical text, our work advances this field by focusing on a unique and specialized annotation process to enhance data quality for these challenging diagnoses.
- We developed and evaluated three NLP-based classification models for CHD (SVM, XGBoost, and RoBERTa) and two for ASD (SVM and RoBERTa),

showcasing the feasibility and effectiveness of using clinical notes in accurately identifying cases of CHD and ASD within a cohort with CHD. Our findings underscore the potential of these models to support accurate clinical decision-making based on narrative medical records, an area that remains underexplored for these specific conditions.

- A rigorous analysis of potential limitations of SVM, XGBoost, and RoBERTa models for CHD and ASD classification revealed that combining linguistic and ML features might improve overall model performance.

5 | Limitations

One limitation of the current study is that ambiguous documentation in the medical record may limit model performance. Cases may have documentation of both ASD and PFO in different clinical text notes, reflecting the difficulty physicians may have in clinically distinguishing between ASD and PFO, particularly physicians who are not specialists in CHD. During the case validation process, abstractors assigned a diagnosis based on a thorough review of medical records, prioritizing the diagnosis of congenital specialists in making a determination. Future studies could similarly prioritize inclusion of specialist notes or imaging notes in NLP models. Moreover, the model's performance can be constrained by the imbalanced distribution between individuals with and without CHD. The CHD classification model was trained on a data set predominantly composed of individuals with CHD. This imbalance might compromise the model's robustness when encountering data sets involving a larger proportion without CHD. The potential application of the current study's methodology may also be constrained by the necessity for clinical text notes to be available in eHRs. Cases that lack clinical text notes could not be included in the current study's cohort. The absence of clinical text notes could be due to limitations of the legacy medical record systems where notes must be scanned into a patient's eHR, and thus not accessible for NLP methods, or due to outdated medical records systems that do not have access to older clinical text notes at all. Nevertheless, now, as almost all health systems are digitized, the use of NLP-based methods for cohort creation will prove highly useful going forward beyond the scope of this study. This study is strengthened by the large, validated data set composed primarily of CHD cases for training the supervised learning models. Replication of our findings in additional data sets is needed.

6 | Conclusions

This study provides important insights into the performance of SVM, RoBERTa, and XGBoost models in identifying cases with CHD and cases with ASD for individuals with CHD. The SVM model for both CHD and ASD classification was the top-performing model, but there is potential for combining different models to improve classification performance. Current findings underline the potential of applying NLP techniques to disease classification tasks and also call for an effective approach to combine NLP models and nonlinguistic, feature-based ML models. Incorporating ML and NLP in CHD surveillance may improve accuracy of CHD identification and ultimately improve patient outcomes. The proposed two-step methodology may also be replicated for other patient cohorts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by National Center for Advancing Translational Sciences under Award Number UL1TR002378 and the Medical Informatics and Artificial Intelligence Core (MIAI), which is supported by the Department of Biomedical Informatics, Emory University School of Medicine. This was also supported by the Centers for Disease Control and Prevention National Center on Birth Defects And Developmental Disabilities (NCBDD) under Grant/Award Number: DD19-1902B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding:

This work was supported by the Centers for Disease Control and Prevention National Center on Birth Defects And Developmental Disabilities (NCBDD) (Grant/Award DD19-1902B) and National Center for Advancing Translational Sciences (Award Number UL1TR002378) and the Medical Informatics and Artificial Intelligence Core (MIAI), which is supported by the Department of Biomedical Informatics, Emory University School of Medicine.

Data Availability Statement

Research data are not shared.

References

- Agarwal A, Thombley R, Broberg CS, et al. 2019. "Age- and Lesion-Related Comorbidity Burden Among US Adults With Congenital Heart Disease: A Population-Based Study." *Journal of the American Heart Association* 8, no. 20: e013450. 10.1161/JAHA.119.013450. [PubMed: 31575318]
- Burchill LJ, Gao L, Kovacs AH, et al. 2018. "Hospitalization Trends and Health Resource Use for Adult Congenital Heart Disease-Related Heart Failure." *Journal of the American Heart Association* 7: e008775. <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L623487677%0A10.1161/JAHA.118.008775>. [PubMed: 30371225]
- Banerjee I, Davis MA, Vey BL, et al. 2023. "Natural Language Processing Model for Identifying Critical Findings—A Multi-Institutional Study." *Journal of Digital Imaging* 36: 105–113. [PubMed: 36344632]
- Broberg C, McLarry J, Mitchell J, et al. 2015. "Accuracy of Administrative Data for Detection and Categorization of Adult Congenital Heart Disease Patients From an Electronic Medical Record." *Pediatric Cardiology* 36: 719–725. [PubMed: 25428778]
- Chang C-C, and Lin C-J. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology* 2, no. 3: 1–27. 10.1145/1961189.1961199.
- Chen T, and Guestrin C. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. Association for Computing Machinery.
- Deniwar A, Hernandez J, Aregullin EO, et al. 2023. "Atrial Septal Defect-Associated Pulmonary Hypertension With Decompensated Heart Failure: Outcomes After Fenestrated Device Closure." *Cardiology in the Young* 34, no. 2: 395–400. 10.1017/S104795112300152X. [PubMed: 37466020]
- Downing KF, Oster ME, Olivari BS, and Farr SL. 2022. "Early-Onset Dementia Among Privately-Insured Adults With and Without Congenital Heart Defects in the United States, 2015–2017." *International Journal of Cardiology* 358: 34–38. [PubMed: 35417738]
- Efron B 1979. "Another Look at the Jackknife." *Annals of Statistics* 7: 1–26.
- Glidewell J, Book W, Raskind-Hood C, et al. 2018. "Population-Based Surveillance of Congenital Heart Defects Among Adolescents and Adults: Surveillance Methodology." *Birth Defects Research* 110: 1395–1403. [PubMed: 30394691]

- Guo Y, Al-Garadi MA, Book WM, et al. 2023. "Supervised Text Classification System Detects Fontan Patients in Electronic Records With Higher Accuracy Than ICD Codes." *Journal of the American Heart Association* 12: 2003–2023.
- Guo Y, Dong X, Al-Garadi MA, Sarker A, Paris C, and Mollá-Aliod D. 2020. "Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets." In *Proceedings of the Australasian Language Technology Workshop*, edited by Kim M, Beck D, and Mistica M, 86–91. Virtual Workshop: Australasian Language Technology Association.
- Guo Y, Ge Y, Yang Y-C, Al-Garadi MA, and Sarker A. 2022. "Comparison of Pretraining Models and Strategies for Health-Related Social Media Text Classification." *Healthcare* 10: 1478. [PubMed: 36011135]
- Gurvitz M, Dunn JE, Bhatt A, et al. 2020. "Characteristics of Adults With Congenital Heart Defects in the United States." *Journal of the American College of Cardiology* 76: 175–182. [PubMed: 32646567]
- Ivey LC, Rodriguez FH, Shi H, et al. 2023. "Positive Predictive Value of International Classification of Diseases, Ninth Revision, Clinical Modification, and International Classification of Diseases, Tenth Revision, Clinical Modification, Codes for Identification of Congenital Heart Defects." *Journal of the American Heart Association* 12, no. 16: e030821. 10.1161/JAHA.123.030821. [PubMed: 37548168]
- Kauling RM, Pelosi C, Cuypers JAAE, et al. 2024. "Long Term Outcome After Surgical ASD-Closure at Young Age: Longitudinal Follow-up up to 50 Years After Surgery." *International Journal of Cardiology* 397: 131616. 10.1016/j.ijcard.2023.131616. [PubMed: 38030038]
- Kaur R, Ginige JA, and Obst O. 2023. "AI-Based ICD Coding and Classification Approaches Using Discharge Summaries: A Systematic Literature Review." *Expert Systems with Applications* 213: 118997. 10.1016/j.eswa.2022.118997.
- Khalifa A, and Meystre S. 2015. "Adapting Existing Natural Language Processing Resources for Cardiovascular Risk Factors Identification in Clinical Notes." *Journal of Biomedical Informatics* 58: S128–S132. [PubMed: 26318122]
- Khan A, Ramsey K, Ballard C, et al. 2018. "Limited Accuracy of Administrative Data for the Identification and Classification of Adult Congenital Heart Disease." *Journal of the American Heart Association* 7, no. 2: e007378. 10.1161/JAHA.117.007378. [PubMed: 29330259]
- Liu Y, Ott M, Goyal N, et al. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv*. 10.48550/arXiv.1907.11692.
- Lui GK, Sommerhalter K, Xi Y, et al. 2022. "Health Care Usage Among Adolescents With Congenital Heart Defects at 5 Sites in the United States, 2011 to 2013." *Journal of the American Heart Association* 11, no. 18: e026172. 10.1161/JAHA.122.026172. [PubMed: 36102252]
- Masud JHB, Kuo C-C, Yeh C-Y, Yang H-C, and Lin M-C. 2023. "Applying Deep Learning Model to Predict Diagnosis Code of Medical Records." *Diagnostics* 13, no. 13: 2297. [PubMed: 37443689]
- Miotto R, Li L, Kidd BA, and Dudley JT. 2016. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients From the Electronic Health Records." *Scientific Reports* 6: 26094. 10.1038/srep26094. [PubMed: 27185194]
- Rajkomar A, Dean J, and Kohane I. 2019. "Machine Learning in Medicine." *New England Journal of Medicine* 380: 1347–1358. [PubMed: 30943338]
- Rodriguez FH, Ephrem G, Gerardin JF, Raskind-Hood C, Hogue C, and Book W. 2018. "The 745.5 Issue in Code-Based, Adult Congenital Heart Disease Population Studies: Relevance to Current and Future ICD-9-CM and ICD-10-CM Studies." *Congenital Heart Disease* 13: 59–64. [PubMed: 29266726]
- Rodriguez FH, Raskind-Hood CL, Hoffman T, et al. 2022. "How Well Do ICD-9-CM Codes Predict True Congenital Heart Defects? A Centers for Disease Control and Prevention-Based Multisite Validation Project." *Journal of the American Heart Association* 11, no. 15: e024911. 10.1161/JAHA.121.024911. [PubMed: 35862148]
- Shi H, Book W, Raskind-Hood C, et al. 2023. "A Machine Learning Model for Predicting Congenital Heart Defects From Administrative Data." *Birth Defects Research* 115, no. 18: 1693–1707. 10.1002/bdr2.2245. [PubMed: 37681293]

- Stanfill MH, Williams M, Fenton SH, Jenders RA, and Hersh WR. 2010. "A Systematic Literature Review of Automated Clinical Coding and Classification Systems." *Journal of the American Medical Informatics Association* 17: 646–651. [PubMed: 20962126]
- Xie P, and Xing E. 2018. "A Neural Architecture for Automated ICD Coding." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Gurevych I and Miyao Y, 1066–1076. Association for Computational Linguistics.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

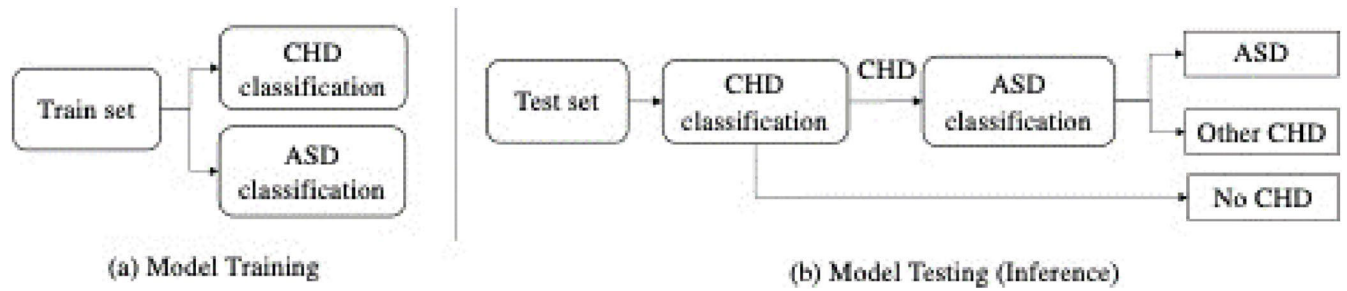


FIGURE 1 |
 System framework for training and testing CHD and ASD classification models. (a) Model training. (b) Model testing (inference). ASD = secundum atrial septal defect; CHD = congenital heart defects.

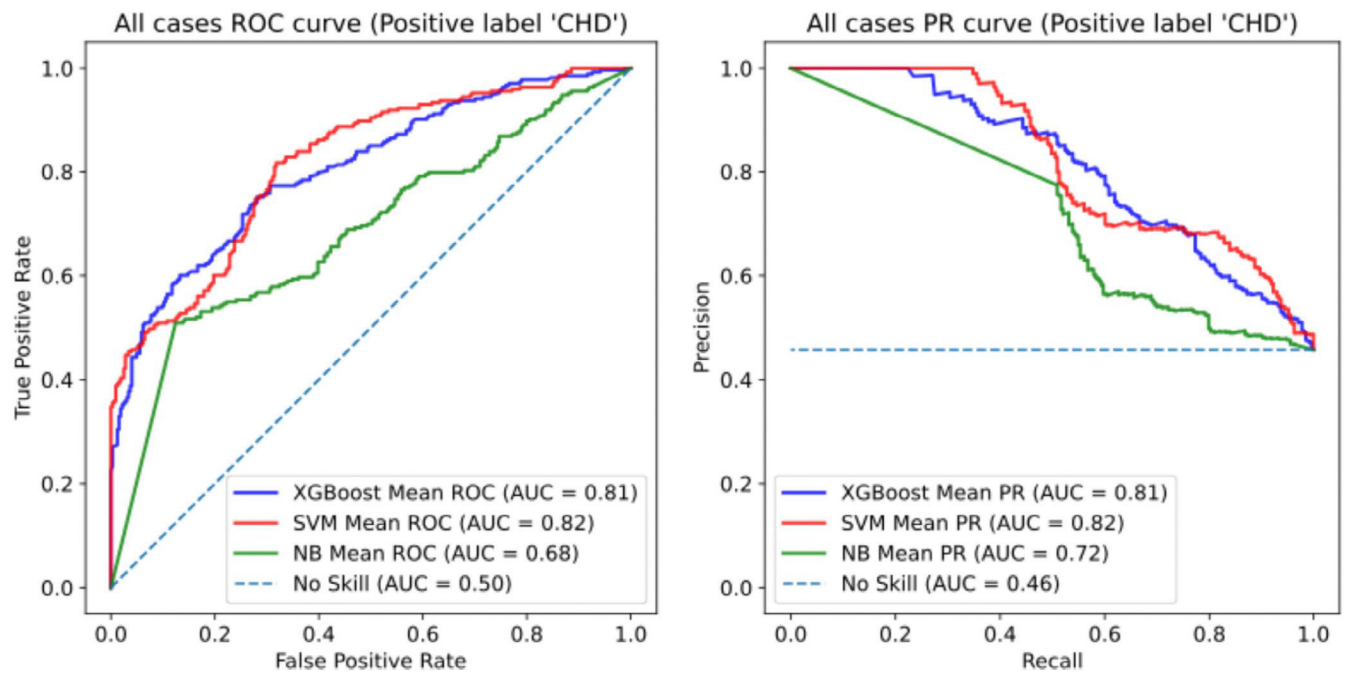


FIGURE 2 |

ROC (left) and PR curves (right) for SVM, XGBoost, NB, and “no skill” models classifying individuals with ASD/PFO ICD codes as having CHD or not, evaluated on the AHS/PHS cohort. For each model, fivefold cross validation was performed. The average estimate and standard deviation over five sub-data sets reported. ASD = secundum atrial septal defect; AUC = area under the curve; CHD = congenital heart defects; ICD = International Classification of Diseases; PFO = patent foramen ovale; PR = precision recall; RoBERTa = a robustly optimized BERT pretraining approach; ROC = receiver operating curve; SVM = Support Vector Machine; XGBoost = a scalable tree boosting system.

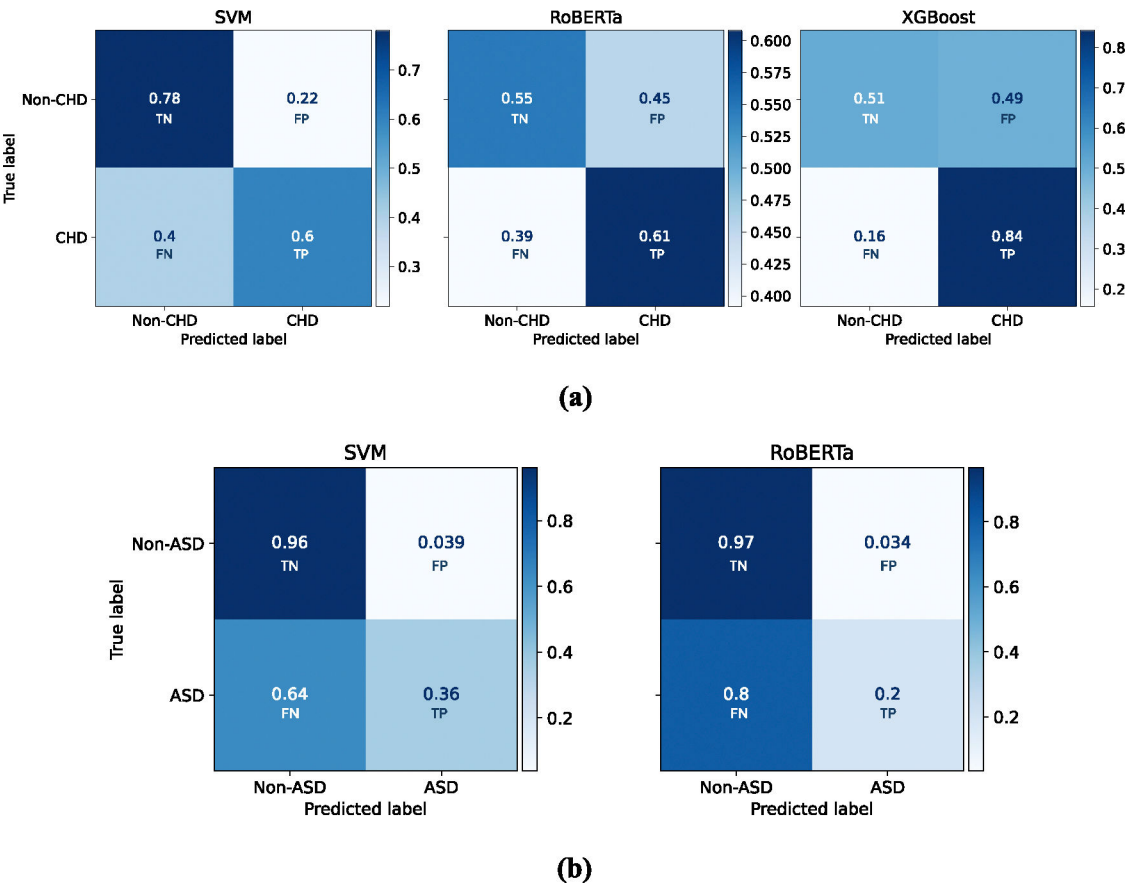


FIGURE 3 | The normalized confusion matrices for models used for CHD and ASD classification, among all cases with ASD/PFO ICD codes identified in the AHS/PHS cohort. (a) Confusion matrices for SVM, RoBERTa, and XGBoost for CHD classification. (b) Confusion matrices for SVM and RoBERTa for ASD classification. ASD = secundum atrial septal defect; CHD = congenital heart defects; FN = false negative; FP = false positive; ICD = International Classification of Diseases; PFO = patent foramen ovale; SVM = support vector machine; TN = true negative; TP = true positive; XGBoost = a scalable tree boosting system.

TABLE 1 |

Cohort characteristics for cases with the ASD/PFO ICD codes (745.5 or Q21.1) and available clinical text notes, 2010–2019.

	AHS/PHS, N = 597 ^a	STS, N = 3256 ^a	<i>p</i> ^b
Mean (SD) Age (1 st encounter)	23.1 (21.6)	3.31 (4.72) ^c	< 0.001
Sex			
Male	279 (46.7%)	1633 (50.2%)	0.12
Female	318 (53.3%)	1623 (49.8%)	
Race			
White	349 (58.5%)	1934 (59.4%)	0.03
Black	190 (31.8%)	1055 (32.4%)	
Other ^d	27 (4.5%)	172 (5.3%)	
Unknown	31 (5.2%)	95 (2.9%)	
Ethnicity			
Hispanic	44 (7.4%)	413 (12.7%)	< 0.001
Non-Hispanic	498 (83.4%)	2770 (85.1%)	
Unknown	55 (9.2%)	73 (2.2%)	
Classified as CHD			
Proportion with a CHD	273 (45.7%)	3223 (99.0%)	< 0.001
PPV (745.5/Q21.1 in isolation) ^e	110/365 (30.1%)	322/340 (94.7%)	< 0.001

Note: Bold *p*-values denote statistically differences between the AHS/PHS and STS groups.

Abbreviations: AHS = Adult Healthcare System; ASD = secundum atrial septal defect; CHD = congenital heart defects; ICD = International Classification of Diseases; PFO = patent foramen ovale; PHS = Pediatric Healthcare System; PPV = positive predictive value; SD = standard deviation; STS = Society for Thoracic Surgeons.

^aSeventy five cases overlap in the two data sets.

^bWilcoxon rank-sum test was used to compare mean age at first encounter between the AHS/PHS cohort and the STS cohort. Sex, race, ethnicity, and other characteristics were defined as in Ivey et al. (2023) and compared using Pearson's chi-square test.

^cA total of 3245 cases had an encounter between 2010 and 2019 used to calculate age at first encounter in 2010–2019.

^dOther racial includes Asian, American Indian/Native American, Native Hawaiian or Other Pacific Islander, and multiracial.

^eIsolated 745.5/Q21.1 = Cases with only ICD code 745.5 or Q21.1, and no other CHD ICD code.

TABLE 2 | System performance of CHD and ASD models for classifying individuals with ASD/PFO ICD codes into three groups, using the data from AHS/PHS.

CHD model	ASD model	Precision	Recall	F_1	95% CI of F_1
AHS/PHS					
<i>Individuals with an ASD</i>					
SVM	RoBERTa	0.90 (± 0.10)	0.18 (± 0.14)	0.28 (± 0.18)	0.21–0.37
SVM	SVM	0.83 (± 0.06)	0.39 (± 0.05)	0.53 (± 0.05)	0.45–0.60
RoBERTa	RoBERTa	0.85 (± 0.19)	0.21 (± 0.16)	0.29 (± 0.18)	0.23–0.38
RoBERTa	SVM	0.78 (± 0.08)	0.40 (± 0.08)	0.52 (± 0.08)	0.44–0.60
XGBoost	RoBERTa	0.85 (± 0.20)	0.19 (± 0.14)	0.28 (± 0.16)	0.22–0.36
XGBoost	SVM	0.77 (± 0.09)	0.38 (± 0.07)	0.51 (± 0.08)	0.43–0.58
<i>Individuals with other CHD</i>					
SVM	RoBERTa	0.24 (± 0.04)	0.35 (± 0.11)	0.28 (± 0.03)	0.21–0.36
SVM	SVM	0.33 (± 0.08)	0.29 (± 0.10)	0.30 (± 0.06)	0.20–0.39
RoBERTa	RoBERTa	0.15 (± 0.05)	0.45 (± 0.21)	0.23 (± 0.08)	0.17–0.28
RoBERTa	SVM	0.16 (± 0.07)	0.38 (± 0.16)	0.22 (± 0.10)	0.15–0.28
XGBoost	RoBERTa	0.24 (± 0.03)	0.91 (± 0.06)	0.38 (± 0.03)	0.32–0.44
XGBoost	SVM	0.25 (± 0.03)	0.84 (± 0.09)	0.39 (± 0.03)	0.33–0.45
<i>Individuals without CHD</i>					
SVM	—	0.69 (± 0.04)	0.92 (± 0.04)	0.78 (± 0.02)	0.75–0.82
RoBERTa	—	0.63 (± 0.07)	0.55 (± 0.07)	0.58 (± 0.06)	0.53–0.63
XGBoost	—	0.79 (± 0.04)	0.51 (± 0.07)	0.62 (± 0.05)	0.57–0.66

Note: The average score and standard deviation over five sub-data sets and 95% confidence interval of F_1 were reported. Bold values represent the highest scores for each performance metric within each model group.

Abbreviations: AHS = Adult Healthcare System; ASD = secundum atrial septal defect; CHD = congenital heart defects; CI = confidence interval; ICD = International Classification of Diseases; PFO = patent foramen ovale; PHS = Pediatric Healthcare System; RoBERTa = a robustly optimized BERT pretraining approach; SVM = support vector machine; XGBoost = a scalable tree boosting system.