Original Article

Clinical SOAP notes completeness checking using machine learning

Sherry J. H. Feng, Jithu Joseph, Edmund M-K Lai

School of Engineering, Computer, and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

Contributions: (I) Conception and design: SJH Feng, EMK Lai; (II) Administrative support: SJH Feng, EMK Lai; (III) Provision of study materials or patients: SJH Feng, J Joseph; (IV) Collection and assembly of data: J Joseph; (V) Data analysis and interpretation: SJH Feng, J Joseph; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Sherry J. H. Feng, PhD. School of Engineering, Computer, and Mathematical Sciences, Auckland University of Technology, 55 Wellesley Street East, Auckland 1010, New Zealand. Email: jiahui.feng@autuni.ac.nz.

Background: Subjective, Objective, Assessment, Plan (SOAP) notes are a critical component of medical documentation. Their completeness and accuracy are critical in healthcare settings. This study evaluates the effectiveness of machine learning methods for analyzing SOAP notes to determine whether each note contains all four required elements.

Methods: Using a dataset of 889 SOAP notes from primary care physician progress notes, we compared multinomial Naive Bayes, logistic regression, random forest, and multilayer perceptron neural networks. We applied standard text preprocessing including lowercasing, tokenization, and vectorization with Term Frequency-Inverse Document Frequency (TF-IDF) weighting. Performance was evaluated using accuracy, precision, recall, and F1 Score through five-fold cross-validation and final testing.

Results: The Naive Bayes classifier outperformed other models, achieving 80.39% accuracy, 81.76% precision, 80.39% recall, and 78.49% F1 Score on the test set. Closer examination revealed that adjusting the probability threshold parameter allows balancing false positives (incorrectly identifying present sections) against false negatives (missing present sections). The superior performance of Naive Bayes suggests that simpler models may be sufficient for SOAP note classification tasks in resource-constrained healthcare settings. The adjustable probability threshold provides flexibility for implementation across different clinical specialties, where documentation practices and tolerance for false positives versus false negatives may vary. Integration into electronic health record systems could enable real-time documentation feedback.

Conclusions: Our findings demonstrate that even simple machine learning approaches can effectively identify missing SOAP elements, potentially improving documentation quality, reducing medical errors, and enhancing communication among healthcare providers. The adaptive algorithm's flexibility allows implementation across different clinical settings with minimal computational resources.

Keywords: Machine learning; clinical documentation; Subjective, Objective, Assessment, Plan notes (SOAP notes); medical informatics; natural language processing

Received: 07 October 2024; Accepted: 10 June 2025; Published online: 20 August 2025.
doi: 10.21037/jmai-24-370
View this article at: https://dx.doi.org/10.21037/jmai-24-370

Introduction

The application of machine learning techniques in the medical domain has been rapidly growing in recent years. Machine learning algorithms have shown remarkable potential for various healthcare tasks, including diagnosis, prognosis, and clinical decision support (1,2). In the context of medical documentation, natural language processing (NLP) techniques can be employed to extract structured information from unstructured clinical notes (3). This can facilitate the automated population of electronic health

records (EHRs), reducing the burden of manual data entry and minimizing errors. Moreover, machine learning algorithms can be trained to identify inconsistencies, gaps, and inaccuracies in medical records, enabling real-time quality control and prompting healthcare providers to rectify documentation issues (4). The integration of machine learning into medical documentation processes can enhance the efficiency of healthcare providers, allowing them to focus on higher-level cognitive work (2). Furthermore, it can promote standardization, ensure completeness, and improve the overall quality of medical records (5).

Subjective, Objective, Assessment, Plan (SOAP) notes

One of the most widely used formats to document essential information collected during a patient visit is the SOAP note (6). A complete note should contain the following four elements:

- Subjective: captures the patient's chief complaint, history of present illness, and relevant past medical history.
- Objective: includes vital signs, physical examination findings, and results of diagnostic tests.
- Assessment: contains the healthcare provider's diagnosis or differential diagnoses based on the subjective and objective data.
- Plan: outlines the treatment strategy, further investigations, and follow-up instructions.

SOAP notes serve as essential tools for communication between healthcare providers, enabling effective continuity of care across different clinicians and settings. Their structured format allows for quick retrieval of critical information and facilitates clinical decision-making. Additionally, SOAP notes serve a crucial medico-legal function, documenting the rationale for clinical decisions and serving as evidence of the standard of care provided. Complete and accurate documentation is essential for proper reimbursement, quality assurance, and medicolegal protection for healthcare providers (7).

With the growing implementation of the OpenNotes movement, which provides patients with direct access to their clinical documentation, the importance of complete and clear SOAP notes has increased dramatically. When patients can view their own medical records, incomplete documentation may lead to confusion, decreased trust in healthcare providers, and potential misunderstandings about their conditions or treatment plans. Complete SOAP notes support transparency in healthcare delivery and can enhance patient engagement in their care (8).

Studies have shown that the quality of SOAP notes varies widely. Common issues encountered include missing or incomplete sections, inconsistencies, and lack of clarity (9). These challenges can be attributed to factors such as time constraints, workload, and the complexity of medical cases (10). Incomplete or inaccurate documentation could lead to communication gaps, medical errors, and suboptimal patient outcomes (11). Therefore, innovative solutions are needed to improve the completeness and accuracy of SOAP notes (12). One way is to automatically check the text of each note to determine if all four elements exist. Naive Bayes is a simple yet powerful algorithm that has been successfully applied in various text classification tasks, including spam email detection (13) and sentiment analysis (14). It is based on the application of Bayes' theorem with strong independence assumptions between the features. Despite its simplicity, Naive Bayes has shown competitive performance

Highlight box

Key findings

- Naive Bayes classifier outperformed other machine learning models for Subjective, Objective, Assessment, Plan (SOAP) note completeness checking, achieving 80.39% accuracy, 81.76% precision, 80.39% recall, and 78.49% F1 Score.
- An adaptive algorithm with adjustable probability threshold was developed that can work with any machine learning model to identify missing SOAP note sections.
- Traditional machine learning approaches performed effectively without requiring extensive computational resources for clinical documentation analysis.

What is known and what is new?

- It is known that incomplete SOAP notes can lead to communication gaps, medical errors, and suboptimal patient care, but manual review of documentation completeness is time-consuming and not scalable.
- This manuscript adds a novel model-agnostic approach for automatically identifying missing SOAP sections, with an adjustable threshold parameter for balancing false positives against false negatives according to clinical needs.

What is the implication, and what should change now?

- Electronic health record systems should incorporate automated SOAP note completeness checking to provide real-time feedback to clinicians.
- Healthcare organizations can implement the proposed adaptive algorithm, customized to their specific documentation patterns and specialties, to improve clinical documentation quality.
- Medical education programs can adopt similar tools to provide immediate feedback during documentation training, establishing proper documentation habits early in clinical careers.

in many domains, making it a potentially viable approach for automated SOAP note checking. In this paper, we examined its performance against three other machine learning models: logistic regression, random forest, and artificial neural networks. In the results section it shows that our SOAP note checking algorithm is the most effective among these machine learning models. In addition, we develop an adaptive algorithm that can be used to check the completeness of SOAP notes. It is flexible enough to work with any machine learning model.

Review of SOAP notes analysis techniques

Assessing the quality and completeness of SOAP notes is a critical task in healthcare organizations. Traditionally, quality assessments of such notes are performed manually by trained professionals, such as physicians, nurses, or medical coders. This involves examining individual SOAP notes to ensure that all required sections are present, the information is accurate and consistent, and the documentation adheres to established guidelines and standards (14). It is a time-consuming and labour-intensive process, especially in large healthcare organizations with a high volume of patient encounters. Hence, the manual review process is not scalable and feedbacks could not be provided in real-time. It is often performed retrospectively on a sample of notes only. Moreover, it is prone to human error and subjectivity, as different reviewers may have varying interpretations of quality criteria and may miss important details (15).

To address the limitations of manual review, researchers have explored automated methods. Rule-based approaches have been developed to extract specific information from SOAP notes using predefined rules and patterns based on linguistic and semantic features (16). These rules are typically created by domain experts and can be effective for targeted information extraction tasks, such as identifying medication information or extracting vital signs. However, developing comprehensive and accurate rules requires extensive domain knowledge and manual effort, which can be time-consuming and resource-intensive. Furthermore, rule-based systems struggle to handle the variability and complexity of clinical language, as they rely on predefined patterns that may not capture all relevant information (17). Such methods also lack the flexibility to adapt to new domains or tasks and require the rules to be manually updated or re-created for each specific application (18).

More recently, machine learning-based approaches have emerged as a promising alternative to rule-based methods for SOAP note analysis and section classification. These approaches leverage supervised learning methods to learn patterns and relationships that can be used to classify SOAP notes into their respective sections or assess their quality. Various machine learning techniques have been applied to SOAP note analysis, including support vector machines (SVM), logistic regression, and random forests (19). These techniques have shown promising results in automatically identifying missing or incomplete sections, detecting inconsistencies, and classifying SOAP notes based on their content. One of the main advantages of machine learning-based approaches over rule-based methods is their ability to learn from large amounts of data and generalize to new examples. Trained models can automatically capture complex patterns and relationships in SOAP notes that may be difficult to express through a set of rules. In addition, they can be more easily adapted to new tasks or domains by retraining the models on relevant labelled data without the need for extensive manual rule design (20). However, machine learning-based approaches for SOAP note analysis also face challenges. Model training requires a large amount of high-quality labelled training data. Annotating SOAP notes with section labels or quality assessments requires significant time and effort from domain experts. The process of manually labelling large datasets is resource-intensive and may also subject to inter-annotator variability (21).

Recent studies in machine learning have further reinforced the importance of structured clinical documentation in improving healthcare outcomes (22). Additionally, Balloch *et al.* (23) examined how AI-assisted documentation tools affected clinician workflow, finding that real-time feedback systems could reduce documentation time by improving completeness.

In recent years, large language models (LLMs) have emerged as a powerful tool for various NLP tasks, including clinical note analysis. LLMs, such as BERT (24), GPT (25), and their variants, are deep neural networks pre-trained on massive amounts of text data. These models capture rich linguistic patterns and semantic relationships, enabling them to generate contextualized word representations. They are often fine-tuned for specific downstream tasks. For example, BioBERT (26), ClinicalBERT (27) and Med-BERT (28) are fine-tuned in large-scale biomedical and clinical corpora, allowing them to better capture the nuances and intricacies of medical language. These domain-specific BERT models have consistently outperformed their general-domain counterparts, highlighting the importance of incorporating domain knowledge when applying LLMs.

Despite the impressive performance of LLMs and domain-specific BERT models in clinical NLP tasks, there are two main concerns:

Interpretability: the deep neural network architectures of LLMs make it difficult to understand how they arrive at their predictions, which can be a concern in healthcare settings where transparency and accountability are crucial (29). Efforts have been made to develop explainable AI techniques for LLMs, but further research is needed to ensure their reliability and trustworthiness in clinical applications (29).

Computational resources: these models have millions or even billions of parameters, making them computationally expensive to train and deploy. Fine-tuning LLMs on domain-specific datasets can also be time-consuming and resource-intensive, which may be a barrier for adoption in resource-constrained settings (30).

Previous efforts have been made to automate SOAP classification in clinical notes. Li *et al.* pioneered work in this area using SVM to classify sections within clinical documents, achieving moderate success with traditional features (31). Similarly, Denny *et al.* developed SecTag, a section header identification system that recognized over 99.5% of section headers in clinical notes but did not specifically focus on the SOAP structure (32). More directly relevant to our work, Deléger *et al.* utilized conditional random fields (CRF) to identify and classify SOAP sections in clinical notes, reporting F-measures in the 90% range for section detection (33).

Our approach differs from these previous efforts in several key aspects. First, while prior work has primarily focused on accurate classification of identified sections, our study emphasizes completeness checking—determining whether all required SOAP sections are present in a clinical note. Second, we introduce an adaptive algorithm with an adjustable probability threshold that can work with any underlying machine learning model, providing greater flexibility than the fixed-model approaches in previous studies. Third, our work specifically evaluates the comparative performance of multiple traditional machine learning techniques for this task, providing practical insights for implementation in resource-constrained healthcare settings. These distinctions highlight the novel contribution of our study to the field of automated clinical documentation analysis.

Research scope and objectives

In this study, we focus on evaluating traditional machine learning models such as Naive Bayes, logistic regression, random forest, and multilayer perceptron feedforward neural networks for automated SOAP note section classification. These models, while not as complex as LLMs, offer several critical advantages for healthcare implementation. They provide greater interpretability, allowing clinicians and administrators to understand and trust the classification decisions. This transparency is essential in healthcare settings where accountability for decisions is required. They also require significantly less computational resources and can be trained and deployed more efficiently, making them practical solutions for resource-constrained healthcare settings with limited IT infrastructure. Additionally, these models can be implemented and maintained with fewer specialized technical resources than required for LLMs, enabling broader adoption across different healthcare environments. Moreover, these models have a long history of successful application in various text classification tasks (34), including clinical document classification (35), providing a proven foundation for SOAP note analysis.

By evaluating and comparing the performance of these traditional machine learning models, our study aimed to provide insights into their suitability for automated SOAP note section classification. This comparative analysis will contribute to a more comprehensive understanding of the strengths and limitations of such modelling techniques, guiding future research and development efforts in automated clinical note analysis.

Methods

In this study, we compared four machine learning models: multinomial Naive Bayes, logistic regression, random forest, and multilayer perceptron neural network for the task of SOAP note section classification. These models were selected due to their widespread use in text classification tasks in the medical domain (2,35,36). The implementation of these methods in the scikit-learn library for machine learning is used. The code used for our experiments is available for public access [Algorithm and Experiment (Google Colab)].

Multinomial Naive Bayes is a variant of the Naive Bayes algorithm specifically suited for text classification tasks (37). It assumes that the features (word counts) follow a multinomial distribution. Given a SOAP note with feature vector x = (x_1, x_2, ... , x_n), where x_i represents the count of the i-th word, the predicted section $\hat{y}$ is given by:

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i | y) \quad [1]$$

The class prior probabilities $P(y)$ and the class-conditional probabilities $P(x_i|y)$ are estimated from the training data using maximum likelihood estimation:

$$P(y) = \frac{N_y}{N} \quad [2]$$

$$P(x_i | y) = \mathrm{I}: \frac{count(x_i, y) + \alpha}{x \in V\left(count(x, y) + \alpha\right)} \quad [3]$$

where N_y is the number of SOAP notes in section y, N is the total number of SOAP notes, $count$ (x_i, y) is the count of word x_i in section y, V is the vocabulary (set of all unique words), and α is a smoothing parameter (typically set to 1 for Laplace smoothing).

Logistic regression is a linear model that estimates the probability of a SOAP note belonging to each section based on the input features (38). The model learns a weight vector w and a bias term b to make predictions. Given a feature vector x, the probability of a SOAP note belonging to section y is given by:

$$P(y | \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}} \quad [4]$$

The model parameters are learned by minimizing the logistic loss function using gradient descent.

Random forest is an ensemble model that combines multiple decision trees to make predictions (39). Each decision tree is trained on a random subset of the training data and a random subset of the features. The final prediction is obtained by aggregating the predictions of all the individual trees. The algorithm for training a random forest is as follows:

Algorithm 1 Random forest training

1: for i = 1 to $n_{estimators}$ do
2: Sample a random subset of the training data with replacement
3: Train a decision tree on the sampled data, considering a random subset of features at each split
4: end for

Multilayer perceptron: we implemented a standard multilayer perceptron (MLP) feedforward neural network. Our network consists of an input layer, one hidden layer with 100 neurons, and an output layer. The activation function used in the hidden layer is the rectified linear unit (ReLU). The output of the network for a SOAP note with feature vector x is given by:

$$\mathbf{y} = \mathrm{softmax}\left(\mathbf{W}_2 \left(\mathrm{ReLU}\left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1\right)\right) + \mathbf{b}_2\right) \quad [5]$$

where W_1 and b_1 are the weights and biases of the hidden layer, W_2 and b_2 are the weights and biases of the output layer, and softmax is the softmax activation function. The network parameters are learned using backpropagation and gradient descent.

Dataset

The data set used in this study is the SOAP notes dataset from the Biomegix repository (40). It consists of clinical text segments labeled with their corresponding SOAP sections: Subjective, Objective, Assessment, and Plan. It is important to clarify the structure of this dataset: each row represents a text segment paired with a single label indicating which SOAP section it belongs to (S, O, A, or P). A complete clinical note would typically be represented as multiple rows in the dataset, with different segments of the note labeled according to their appropriate sections.

The original dataset contains 889 labeled text segments. In our preprocessing, we maintained this structure where each segment is associated with a single section label. We performed standard text preprocessing steps including lower-casing, removal of special characters, and tokenization. We did not remove stop words, as these can carry meaningful structural information in clinical text. The data set was divided into a training set of 838 segments and a testing set of 51 segments, maintaining the original distribution of section labels.

The distribution of sections in the dataset is relatively balanced: subjective (26.3%), objective (28.1%), assessment (23.5%), and plan (22.1%). This balance helps ensure that models do not develop strong biases toward any particular section during training.

Statistical analysis

To ensure the robustness of the evaluation, we employed a two-stage approach. First, five-fold cross-validation was performed on the training set of 838 text segments for

model selection and hyperparameter tuning. The training set was divided into five equally sized lots, and each model was trained and evaluated five times. In each iteration, a different lot was used for validation while the other four were used for training.

After finalizing the models based on cross-validation results, we evaluated their performance on the separate held-out test set of 51 text segments. The performance metrics reported in *Table 1* are based on this final evaluation using the test set, providing an unbiased assessment of model performance on unseen data. This approach ensures that our reported results reflect the models' generalization capability rather than their ability to fit the training data.

Four standard metrics: accuracy, precision, recall, and F1 Score were used in this study (41):

- Accuracy: the proportion of correctly classified SOAP notes out of the total number of notes.
- Precision: the proportion of true positive predictions among all positive predictions for each section.
- Recall: the proportion of true positive predictions among all actual positive instances for each section.
- F1 Score: the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

For precision, recall, and F1 Score, we used the weighted average across all SOAP sections to account for class imbalance. All statistical analyses and model implementations were performed using Python with the scikit-learn library. Comparative analysis between the four machine learning models was based on direct comparison of these performance metrics, with differences in accuracy of more than 2 percentage points considered practically significant in the clinical documentation context. To analyze the sensitivity of the SOAP Checker algorithm to different probability threshold settings, we qualitatively examined its performance on sample notes, which helped understand the trade-off between falsely identifying present sections (false positives) and missing present sections (false negatives).

SOAP Checker

In order to gain insights into the behaviour of the machine learning algorithms, a SOAP Checker program is developed. It can work with any algorithm, including the four detailed in this paper, to identify missing SOAP note sections. Algorithm 2 describes the details of the algorithm used in this program.

Algorithm 2 SOAP missing checker

Require: *text*: Input text, *clf*: Classifier, *vec*: Vectorizer, *input_probability*: float

```
Initialize missing ← {S: True, O: True, A: True, P: True}
for each line in split(text,n) do
  if strip(line) ≠ '' then
    input_vector ← vec.transform([line])
    prob ← clf.predict_prob(input_vector)[0]
    label_prob ← sorted(zip(clf.classes_, prob), key = λx : x[1], reverse=True)
    for each label, p in label_prob do
      if p > input_probability then
        missing[label] ← False
      end if
    end for
  end if
end for
result ← ["Missing: " + k for k, v in missing.items() if v]
if len(result) == 0 then
  append(result, "All SOAP sections filled")
end if
return result
```

Results

The performance of the four models in classifying SOAP note sections is presented in *Table 1*. These results show that the Naive Bayes classifier achieved the best performance across all four-evaluation metrics (accuracy: 80.39%, precision: 81.76%, recall: 80.39%, F1 Score: 78.49%). The MLP and random forest models also performed well (78.49% and 78.43% in accuracy respectively), while the Logistic Regression model, while slightly less effective (accuracy: 76.47%), still provided reasonable results. Four instances of the results of applying the SOAP Checker with multinomial Naive Bayes, logistic regression, and random forest are shown in *Table 2*. In all four notes, multinomial Naive Bayes provided accurate identification of missing or present sections. For example, it correctly identified all sections as present in the comprehensive depression note (Note 2) and accurately indicated that only the Plan section was present in the medication list (Note 3). Our results

Table 1 Performance metrics of different models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic regression	76.47	80.82	76.47	74.56
Random forest	78.43	80.74	78.47	76.88
Naive Bayes	80.39	81.76	80.39	78.49
MLP	78.49	81.59	78.43	77.03

MLP, multilayer perceptron.

demonstrate that the probability threshold parameter is a critical factor affecting the performance of the SOAP Checker algorithm. Varying this threshold produced different results in section identification accuracy even when using the same trained model. For instance, changing the threshold altered which specific sections were flagged as missing in the first note in *Table 2*.

Discussion

The superior performance of Naive Bayes aligns with previous studies highlighting its strength in text classification tasks, particularly with sparse and high-dimensional data (42,43). Its effectiveness can be attributed to its assumption of conditional independence among features. While this assumption may not always hold true, it appears to be a reasonable approximation for SOAP note classification, leading to effective results. The strong performance of simpler models like Naive Bayes suggests that complex deep learning approaches may not be necessary for this specific clinical documentation task.

The performance of MLP and random forest models demonstrates their ability to capture complex patterns in unstructured text data (39,44). Even the logistic regression model provided reasonable results, suggesting that linear models can contribute effectively to SOAP note classification tasks. This is particularly important for healthcare settings with limited computational resources.

Algorithmic constraints

A limitation of the SOAP Checker algorithm lies in its reliance on a user-defined probability threshold, which introduces a degree of variability into its performance. This variability underscores a challenge in implementing the algorithm in real-world settings; users must balance the model's inherent performance with an appropriate threshold selection. Setting a threshold too low may result in wrongly indicating that sections are present when they are not. On the other hand, too high a threshold could lead to sections that are present being flagged as missing. Future improvements could focus on developing guidelines for optimal threshold selection based on the specific model used and the type of SOAP note being analyzed, thereby enhancing the algorithm's consistency and reliability across diverse clinical documentation scenarios.

Clinical implementation of the SOAP Checker

A key strength of our SOAP Checker approach lies in its flexibility and model-agnostic design. As highlighted in the introduction and methodology sections, our algorithm can work with any underlying machine learning classifier that provides probability estimates for section classification. This flexibility offers several important advantages for real-world implementation. First, it allows healthcare organizations to select the classification model that best fits their specific documentation patterns, computational resources, and performance requirements. Second, it enables the system to evolve over time as new machine learning approaches emerge, without requiring fundamental redesign of the checking framework. The organization could upgrade the underlying classification model while maintaining the same interface and threshold-based checking mechanism.

This flexibility also facilitates adaptation to different clinical specialties and settings. For example, a primary care clinic might optimize the classifier and thresholds differently than an emergency department due to their distinct documentation patterns and requirements. The model-agnostic design of our SOAP Checker enables these context-specific adaptations while maintaining a consistent framework for completeness assessment. Furthermore, this design approach simplifies integration with existing EHR systems, as the checking mechanism can be implemented as a modular component that interfaces with the existing documentation infrastructure.

Table 2 Sample of results from the SOAP Checker

SOAP notes	Multinomial Naive Bayes	Logistic regression	Random forest
John has been feeling a bit tired and he finds it hard to eat. He also not feeling happy. Prescribed to get more time outside and take lots of water. John will develop the ability to recognize and manage his depression	Missing: objective, assessment, plan	Missing: objective	Missing: assessment, plan
John was unable to come into the practice and so has been seen at home. John's personal hygiene does not appear to be intact; he was unshaven and dressed in track pants and a hooded jumper which is unusual as he typically takes excellent care in his appearance. John appears to be tired; he is pale in complexion and has large circles under his eyes. John's compliance with his new medication is good, and he appears to have retained his food intake. Weight is stable and unchanged. Assessment John presented this morning with low mood and affect. John exhibited speech that was slowed in rate, reduced in volume. His articulation was coherent, and his language skills were intact. His body posture and effect conveyed a depressed mood. John's facial expression and demeanor were of someone who is experiencing major depression. Affect is appropriate and congruent with mood. There are no visible signs of delusions, bizarre behaviors, hallucinations, or any other symptoms of psychotic process. Associations are intact, thinking is logical, and thought content appears to be congruent. Suicidal ideation is denied. Short and long-term memory is intact, as is the ability to abstract and do arithmetic calculations. Insight and judgment are good. No sign of substance use was present. Plan diagnoses: the diagnoses are based on available information and may change as additional information becomes available. Major depressive disorder, recurrent, severe F33.1 (ICD-10) Active Link to treatment Plan Problem: Depressed Mood Problem: Depressed Mood John's depressed mood has been identified as an active problem requiring ongoing treatment. It is primarily evident through a diagnosis of Major Depressive Disorder. Long-term goal: John will develop the ability to recognize and manage his depression. Short-term goals and interventions: Continue to attend weekly sessions with myself continue to titrate up SSRI, fluoxetine to walk Jingo once a day to use a safety plan if required	All filled	Missing: objective	Missing: plan
Pentamidine 300 mg IV q. 36 hours pentamidine nasal wash 60 mg per 6 mL of sterile water q.d. voriconazole 200 mg p.o. b.i.d. acyclovir 400 mg p.o. b.i.d. cyclosporine 50 mg p.o. b.i.d. prednisone 60 mg p.o. q.d. GCSF 480 mcg IV q.d. epogen 40,000 units subcu q. week protonix 40 mg q.d. simethicone 80 mg p.o. q. 8 nitroglycerin paste 1 &quot q. 4 h. p.r.n. flunisolide nasal inhaler 2 puffs q.8 oxycodone 10–15 mg p.o. q. 6 p.r.n. sudafed 30 mg q. 6 p.o. p.r.n. fluconazole 2 cream b.i.d. to erythematous skin lesions ditropan 5 mg p.o. b.i.d. Tylenol 650 mg p.o. q. 4 h. p.r.n. Ambien 5–10 mg p.o. q. h.s. p.r.n. Neurontin 100 mg q. a.m. 200 mg q. p.m. Aquaphor cream b.i.d. p.r.n. Lotrimin 1 cream b.i.d. to feet Dulcolax 5–10 mg p.o. q.d. p.r.n. Phoslo 667 mg p.o. t.i.d. Peridex 0.12 15 mL p.o. b.i.d. mouthwash Benadryl 25–50 mg q. 4–6 h. p.r.n. pruritus Sarna cream q.d. p.r.n. pruritus Nystatin 5 mL p.o. q.i.d. swish and ! spit folate 1 mg p.o. q.d. vitamin E 400 units p.o. q.d. Haldol 2 mg IV q. 6 p.r.n. agitation Colace 100 mg b.i.d. Senna 2 tablets p.o. b.i.d.	Missing: subjective, objective, assessment	Missing: subjective, objective, assessment	Missing: subjective, objective, assessment
Severe headache. Prescribed paracetamol	Missing: objective, assessment	Missing: objective, assessment, plan	Missing: objective, assessment, plan

IV, intravenous; SOAP, Subjective, Objective, Assessment, Plan; SSRI, Selective Serotonin Reuptake Inhibitor.

For the SOAP Checker to be effectively integrated into clinical workflows, several implementation pathways can be considered. The most promising approach would be integration directly into EHR systems as a real-time feedback tool. In this scenario, the SOAP Checker could analyze clinical notes as they are being written, providing immediate feedback to clinicians about missing sections before the note is finalized. This real-time integration would allow for on-the-spot corrections, potentially improving documentation quality at the point of care.

The tool could also be deployed as a quality improvement instrument for retrospective analysis across departments or individual providers. Healthcare organizations could use it to identify documentation patterns, recognize educational opportunities, and track improvements over time. For instance, department heads could receive regular reports highlighting the completeness rates of SOAP notes within their units, enabling targeted interventions where documentation gaps are prevalent.

Practical considerations for implementation include the need for specialty-specific customization. Different medical specialties may have varying documentation requirements and styles, necessitating adjustments to the probability threshold parameter. For example, emergency medicine notes might require different threshold settings compared to psychiatric evaluations due to the distinct nature of these clinical encounters. Healthcare organizations implementing the SOAP Checker should consider a calibration phase where the tool is fine-tuned for different clinical contexts. This threshold adjustment represents calibration within a consistent algorithmic framework rather than fundamental modification, allowing the system to adapt to specialty-specific documentation patterns while maintaining the core algorithm's integrity.

The SOAP Checker also holds significant potential as an educational tool. Medical schools and residency programs could incorporate it into documentation training, allowing students and residents to receive immediate feedback on their notes. This application could help develop proper documentation habits early in clinical careers, potentially leading to long-term improvements in medical record quality.

User interface considerations are equally important for successful adoption. The feedback provided by the SOAP Checker should be non-intrusive yet clear, without disrupting the clinician's workflow. Integration should be seamless, requiring minimal additional clicks or attention from already time-constrained healthcare providers.

Limitations and future directions

A significant limitation of this study lies in the homogeneity of documentation styles represented in our dataset. Our sample primarily contains narrative, prose-style SOAP notes, which may not adequately represent the diversity of documentation approaches used across different healthcare systems and specialties. Many clinicians use point-form, templated, or hybrid documentation styles that differ substantially from the prose examples in our dataset. This limitation potentially affects the generalizability of our models to real-world clinical settings where documentation practices vary widely.

The performance metrics achieved in our study, while promising, still fall short of human-level accuracy. The highest accuracy of 80.39% achieved by the Naive Bayes classifier would need significant improvement before widespread clinical implementation could be considered. For a system to be viable in clinical settings, where documentation errors can impact patient care, we estimate that accuracy levels of at least 95% would be necessary.

To address both the performance and generalizability limitations, several research directions warrant exploration:

- Diverse datasets: incorporating data from EHR databases such as MIMIC-III or MIMIC-IV would provide access to a wider range of clinical documentation styles from various care settings. Training models on datasets that encompass different documentation approaches—including point-form notes, semi-structured templates, and specialty-specific formats—would likely improve real-world applicability.
- Specialty-specific adaptations: different medical specialties may emphasize different aspects of the SOAP framework or use specialty-specific terminologies. Future research could explore specialty-calibrated models or adaptive approaches that accommodate these variations.
- Advanced feature engineering: more sophisticated approaches could enhance classification accuracy, including word embeddings that capture semantic relationships between medical terms, domain-specific medical ontologies like Unified Medical Language System (UMLS) to provide structured medical knowledge, and n-gram features to better capture phrasal patterns characteristic of different SOAP sections.
- Ensemble and hybrid methods: combining multiple

classifiers through stacking or voting systems might yield better results than any single classifier. For example, integrating Naive Bayes with selective use of transformer-based models could balance computational efficiency with improved accuracy for challenging cases. Additionally, incorporating rule-based heuristics to leverage section headers or specific terminology could improve classification in ambiguous cases.

- Class imbalance techniques: specialized approaches for handling imbalanced data could further enhance model performance, particularly for sections that appear less frequently in typical notes. Methods such as Synthetic Minority Oversampling Technique (SMOTH) (45) or class-weighted learning deserve investigation.

The strategies outlined above, particularly when combined, offer promising pathways to achieving the performance threshold required for clinical implementation while ensuring generalizability across diverse healthcare documentation practices. Future iterations of the SOAP Checker should prioritize both performance improvement and adaptability to various documentation styles to maximize clinical utility.

Conclusions

A comparative analysis of four different machine learning methods—multinomial Naive Bayes, logistic regression, random forest, and multilayer perceptron, has been made for the identification of presence of the four sections of a SOAP note. We found the Naive Bayes classifier outperformed the other three. In closer examination of the behaviour of this classifier, we note that its performance is dependent on a probability threshold. Hence, it will require some fine-tuning to obtain a balance between false positives and false negatives.

Future research directions could include the exploration of ensemble methods that combine the strengths of multiple models, such as stacking or voting, to further improve the classification performance. Moreover, incorporating domain-specific knowledge, such as medical ontologies or clinical guidelines, into the feature engineering process could potentially enhance the model's ability to capture relevant information and improve their classification accuracy.

Acknowledgments

None.

Footnote

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-370/prf

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-24-370/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. No Insitutional Review Board (IRB) approval or informed consent is required.

References

1. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24-9.
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230-43.
3. Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016;23:1007-15.
4. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag 2013;10:1c.
5. Raghu M, Blumer K, Sayres R, et al. Direct uncertainty prediction for medical second opinions. In: International Conference on Machine Learning. PMLR, 2019;97:5281-90.
6. Cameron S, Turtle-Song I. Learning to write case

notes using the soap format. Journal of Counseling & Development 2002;80:286-292.
7. Weed LL. Quality control and the medical record. Arch Intern Med 1971;127:101-5.
8. Delbanco T, Walker J, Bell SK, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med. 2012;157:461-70.
9. Edwards ST, Neri PM, Volk LA, et al. Association of note quality and quality of care: a cross-sectional study. BMJ Qual Saf 2014;23:406-13.
10. O'Donnell HC, Kaushal R, Barrón Y, et al. Physicians' attitudes towards copy and pasting in electronic note writing. J Gen Intern Med 2009;24:63-8.
11. Stetson PD, Morrison FP, Bakken S, et al. Preliminary development of the physician documentation quality instrument. J Am Med Inform Assoc 2008;15:534-41.
12. Androutsopoulos I, Koutsias J, Chandrinos KV, et al. An evaluation of naive bayesian anti-spam filtering. arXiv preprint cs/0006013 2000.
13. Liu B. Sentiment analysis and opinion mining. Springer Nature; 2012.
14. Stetson PD, Bakken S, Wrenn JO, et al. Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3:164-74.
15. Hammond KW, Helbig ST, Benson CC, et al. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. AMIA Annu Symp Proc 2003;2003:269-73.
16. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;128-44.
17. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
18. Friedman JH, Popescu BE. Predictive learning via rule ensembles. The Annals of Applied Statistics 2008;2:916-54.
19. Mujtaba G, Shuib L, Idris N, et al. Clinical text classification research trends: systematic literature review and open issues. Expert Systems with Applications 2019;116:494-520.
20. Pustejovsky J. Iso-timeml and the annotation of temporal information. In: Ide N, Pustejovsky J. (eds). Handbook of Linguistic Annotation. Springer, Dordrecht; 2017:941-68.
21. Roberts K, Demner-Fushman D, Voorhees EM, et al. Overview of the TREC 2017 Precision Medicine Track. Text Retr Conf 2017;26:https://trec.nist.gov/pubs/trec26/papers/Overview-PM.pdf.
22. Perkins SW, Muste JC, Alam T, et al. Improving Clinical Documentation with Artificial Intelligence: A Systematic Review. Perspect Health Inf Manag 2024;21:1d.
23. Balloch J, Sridharan S, Oldham G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthc J 2024;11:100157.
24. Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018.
25. Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018.
26. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234-40.
27. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 2019.
28. Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4:86.
29. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel) 2020;23:18.
30. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI Conference on Artificial Intelligence 2020;34:13693-6.
31. Li Y, Lipsky Gorman S, Elhadad N. Section classification in clinical notes using supervised hidden markov model. In: Proceedings of the 1st ACM International Health Informatics Symposium. 2010:744-750.
32. Denny JC, Spickard A 3rd, Johnson KB, et al. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc 2009;16:806-15.
33. Deléger L, Grouin C, Zweigenbaum P. Extracting medication information from French clinical texts. Stud Health Technol Inform 2010;160:949-53.
34. Aggarwal CC, Zhai C. A survey of text classification algorithms. In: Aggarwal, C., Zhai, C. (eds). Mining Text Data. Springer, 2012:163-222.
35. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352-9.
36. Liaw A, Wiener M. Classification and regression by randomforest. R news 2002;2:18-22.
37. Manning CD. Introduction to information retrieval.

Syngress Publishing, 2008.
38. Hastie T, Tibshirani R, Friedman JH, et al. The elements of statistical learning: data mining, inference, and prediction. NY: Springer; 2009.
39. Breiman L. Random forests. Machine learning 2001;45:5-32.
40. Biomegix. Soap notes dataset 2020, Available online: https://huggingface.co/datasets/biomegix/soap-notes
41. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 2009;45:427-37.
42. Zhang H. The Optimality of Naive Bayes. Proceedings of 17th International Florida Artificial Intelligence Research Society Conference, Menlo Park, 12-14 May 2004, 562-7.
43. Kim SB, Han KS, Rim HC, et al. Some effective techniques for naive bayes text classification. IEEE Transactions on Knowledge and Data Engineering 2006;18:1457-66.
44. Goldberg Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research 2016;57:345-420.
45. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002;16:321-57.

doi: 10.21037/jmai-24-370
Cite this article as: Feng SJH, Joseph J, Lai EMK. Clinical SOAP notes completeness checking using machine learning. J Med Artif Intell 2025.