SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Early Detection of Cardiovascular Disease Through Symptom-Based Text Mining: A Pure Natural Language Processing Approach Using the MIMIC-III Public Dataset

Pavitra Rachala[1], B. Padmaja Rani[2]

[1]Research scholar, Department of CSE, JNTUH & Assistant Professor, Department of CSE(CS), CVR College of Engineering, Hyderabad, India
[2]Senior professor, Dept of CSE, JNTUHUCES, Kukatpally, Hyderabad, Telangana, India pavithra.rachala@gmail.com

Abstract

Despite being reduced, cardiovascular disease (CVD) is still the leading cause of death in the world, and misrecognition of early warning signs, which includes chest discomfort, shortness of breath, palpitations and lack of energy, persist leading to poor outcomes. Although there are techniques for imaging and biomarker-based screening, these methods are expensive, unavailable for many, and the first and most common source of information about a patient's condition is the free-text clinical narrative in primary-care settings and in low-resource areas. This paper suggests a novel image-free,text-based, non-invasive method for early detection of the Cardiovascular Diseases (CVD) based entirely on Natural Language Processing (NLP) methods used on the unstructured Clinical Notes (CNs). We use the publicly available MIMIC-III critical care database to model a two-stage pipeline, a Symptom Extraction and Weighting (SEW) algorithm that extracts and weights cardiovascular symptom mentions from free text using a lexicon-based named entity recognition approach and Term Frequency-Inverse Document Frequency (TF-IDF) statistics, and a Risk Scoring and Classification (RSC) algorithm that combines the weighted symptom evidence in free text via a guideline-based clinical weight table, and uses a sigmoid-based decision function to output a meaningful Low/High/Moderate CVD risk class. Both algorithms are intentionally formulated in a simple closed-form mathematical way to enforce a low-calculability cost, to make the pipeline fully interpretable and without the need for GPU infrastructure for deployment. The Proposed pipeline achieved competitive accuracy of 87.6%, an AUROC of 0.889, and an F1-score of 0.85, and has significantly lower inference time and memory usage, not only when compared to the considerably heavier transformer-based baselines, but also and most notably, when compared to the results on the held-out test set of 8,432 discharge summaries and progress notes on MIMIC-III. The results suggest that the well-designed and purely statistical/rule based NLP pipelines are still worthwhile and explainable for early and low-cost, symptom-based cardiovascular risk screening in practice.

Keyword: Cardiovascular disease, early detection, symptom extraction, natural language processing, TF-IDF, MIMIC-III, clinical text mining

Introduction

Cardiovascular disease (CVD) is responsible for about 18 million deaths every year and the greatest single cause of burden of non-communicable disease (NCDs) around the world. Much of this responsibility is because of not only lack of symptoms, but due to the lack of recognizing, documenting and acting on symptom patterns at a sufficiently early stage to intervene. Structured variables that are offered by traditional risk-stratification tools like the Framingham Risk Score and the ASCVD Pooled Cohort Equations are manually abstracted from the chart or cannot easily be found at first contact. Since the early symptomatic history (complaints of chest tightness, shortness of breath, palpitations, or fatigue) is what we look back on and use to develop a good practice of anthropology, the majority of it is recorded as unstructured free text in nursing notes, triage notes, and progress notes made by physicians. Natural Language Processing (NLP) provides a direct way to capture this unstructured "signal" without the need for imaging, lab tests, extra data collection on clinicians or anything else. A purely text-based detection pipeline is appealing for three practical reasons: (1) there is no added marginal cost to the data collection because the information is already in the electronic health record (EHR), (2) it can be deployed on existing records backward-looking to find cases, and (3) it can be constructed from clear, transparent, and explainable statistical components instead of opaque deep network architectures, allowing for an EHR output that is clinically interpretable and that can be audited by a physician. While the field has seen significant progress with deep learning and large language models (LLMs) applied to clinical NLP, from pioneering deep architectures for

tagging risk factors, to contemporary transformer-based EHR models, there continues to be a technical challenge: most currently advanced symptom-extraction systems require a considerable amount of compute, annotated training data, and dedicated hardware — resources that are hard to find in many healthcare institutions. This paper tackles this problem by focusing the question: Can high quality early detection performance be reached with simple mathematically transparent NLP-features such as lexicon based symptom recognition, TF-IDF weighting, and rule drawing without having to rely on deep neural architectures is major motivated question on this work. The main contributions of this paper are as follows:

We are introducing and using an early CVD detection pipeline (no imaging, no structured EHR field) that is entirely based on publicly available MIMIC-III database to ensure 100% reproducibility of our workflow/dataset with other research groups. We developed two algorithms: (2) Symptom Extraction and Weighting (SEW): The algorithm is segmented into five elementary steps, which contains basic mathematical tasks (frequency counting, weighted sum, normalization, applying a sigmoid function) alongside elementary NLP tasks (entity matching by dictionary and negation handling). (3) Risk Scoring and Classification (RSC): The algorithm is broken down into five simple steps that include elementary mathematics (frequency counting, weighted sum, normalization, applying sigmoid function) and basic NLP (tokenization, handling negation, entity matching). Towards achieving the aforementioned, we extensively compare our lightweight pipeline with the top three transformer-based alternatives by rate, algorithmic complexity, and tabular comparisons, and validate that the proposed pipeline attains competitive detection performance at a significantly lower cost in terms of computational resources. We measure for the first time in this particular symptom-detection application, the (sacred) cow of interpretability/deployment cost vs. marginal gain in terms of accuracy, for the two NLP approaches: purely statistical vs. deep learning/LLM.

The rest of this paper is organized as follows. The review of previous research on clinical NLP for cardiovascular disease within the last eight years (from 2018 to 2025) is found in Section 2. The MIMIC-III derived dataset and the proposed methodology (architecture diagram and two algorithms) are described in section 3. Section 4 presents results, consolidated in four tables, that first contain the statistics on datasets, second the complexity of the algorithms, and third give comparison with state-of-the-art approaches. The implications and limitations of the findings have been discussed in Section 5 while Section 6 will conclude the paper.

Related Work

There is extensive evolution of Clinical NLP for cardiovascular performance over the seven last years, shifting from rule based information extraction to pipelines enhanced by deep learning to the newest pipelines augmented by transformer/LLMs. It is worth to note and the purpose of this section is to review some of the representative work carried out in the time period 2018-2025 in a chronological order to show how this path was followed.

One of the earlier studies that welcome a discussion about the comparative success of deep learning versus hybrid rule-and-dictionary systems for heart disease risk factor extraction was performed by Chokwijitkul et al. (2018), where they tried several different neural architectures within the 2014 i2b2/UTHealth clinical narrative corpus and conclude that data-driven deep models performance may be comparable or superior, but significantly more expensive in terms of annotated data, than the hybrid rule-and-dictionary systems [1].

Sheikhalishahi et al. [2] conducted an early systematic review of the application of NLP to clinical notes with chronic conditions, mapping 102 discrete chronic conditions found in the literature and describing persistent on-going problems of using NLP to shift from basic entity extraction towards more holistic comprehension, temporal reasoning, and relation extraction between the clinical entities. Devlin et al. (2019) released BERT, a bidirectional transformer architecture that would be to form the basis of almost all state-of-the-art clinical NLP systems explored in the comparison tables throughout this paper.

A recent study from 2020 [4] that was published in the journal Healthcare builds upon the classical feature engineering and proposes a hybrid feature representation technique for multiclass classification of cardiovascular biomedical texts that combines bag-of-words and word-embedding features with statistical weighting strategies – term frequency, inverse document frequency, and class probability – to enhance multiclass cardiovascular text discrimination. It employs a similar method to TF-IDF-based weighting that is also used in the current paper, but differs in that it addresses document-level topic classification instead of symptom-level extraction.

This issue is directly addressed in Zhan et al. (2021), who make a comparison between the older and newer techniques—TF-IDF vs. Word2Vec, Doc2Vec—plus "logistic regression" across two Stanford cohorts, and, crucially, transferability to the publicly available MIMIC-III [5] database. Their conclusion that interpretable TF-

IDF-based models matched or surpassed less-interpretable embedding models directly inspired the design choice in this paper of weighting algorithms that are transparent and easy to compute over embedding weights that are quite opaque and difficult to compute. Berman et al. (2021) later creates five rule-informed NLP modules to identify hypertension, dyslipidaemia, diabetes, coronary artery disease and stroke/transient ischaemic attack in free-text notes from Mass General Brigham, with a sensitivity, specificity and PPV all consistently over 85% [6].

In the field of cardiology, that is the condition with which this review will be focused, a systematic review of NLP methods and applications was published in 2022 (Turchioe et al. [7]). This review synthesized 37 studies and covered heart failure as well as other conditions such as coronary artery disease, electrophysiology and general cardiology. One of the largest drawbacks identified is the lack of uniformity among the studies in the field; this is particularly evident in how NLP methods, assessment procedures, and reporting of results are employed. The exclusion of results from aggregation makes for a major motivation for the equipping of standardized complexity and performance comparison tables as presented in Section 4 of this paper. In the following year Rao et al. (2022) suggested an explainable transformer for predicting incident heart failure from longitudinal linked electronic health records, covering over 100,000 patients in the United Kingdom and demonstrated the potential for attention-based explainability methods to partly overcome the lack of interpretability in deep clinical NLP models [8].

As part of the 2014 i2b2/UTHealth challenge track, Houssein, Mohamed, and Ali (2023) showed impressive results on detecting heart disease risk factors (diabetes, coronary artery disease, hyperlipidemia, hypertension, smoking, obesity) from clinical notes, usually reported as sentences, and, to their credit, demonstrated that stacking of word embeddings (a form of embedding) was superior to previous hybrid embedding (rule) and deep-learning systems on this task [9]. The present study uses this paper as a measure of the performance of deep-learning models (Table 3). In the same year, Johnson et al. (2023) released an updated and expanded version of MIMIC-III called MIMIC-IV, which increased the number of publicly available EHRs for cardiovascular NLP research and provided a new extension corpus for the purposes of future validation of the pipeline presented here [10].

The authors of this paper, Deepa et al. (2024), surveyed and used machine learning methods for the early prediction of cardiovascular diseases from health records, and found that in health records, unstructured clinical notes or occurrence of symptoms such as "fatigue" contain important early qualitative signals that are not routinely captured by structured risk scores such as the Framingham risk score or ASCVD score. Their focus on early qualitative symptom signal is directly connected to the motivation of using a symptom-centric (rather than solely a biomarker-centric) detection approach.

Recently, Diaz Ochoa et al. (2025) compared the performance of a zero-shot large language model approach (mistral-neo-instruct, GOTT) with a fine-tuned BioGottBERT model for recognition of symptoms, negation and anatomy in emergency department narratives, revealing that our data-protected fine-tuned transformer model outperformed the zero-shot LLM approaches on the test set (F1 = 0.84) and required an on-premises and data-protection-compliant deployment [12]. This paper presents the current status of the state-of-the-art (SOTA) transformer baseline that is benchmarked in comparison with the proposed lightweight pipeline in Section 4. In the same year, using a different weight method, TF IDF, in the context of rheumatology, TF-IDF weighted free-text symptom descriptions directly written by the patients fulfilled diagnostic decision making in a clinically relevant fashion, further underscoring that classical and comprehensible statistical weighting methods have also staying power beyond the context of cardiology [13].

In each piece of this work, a trade-off between the detection capability and computational/interpretability cost arises as well – transformer and LLM-based methods are often marginally better in terms of their F1-score, but significantly more costly to run and deploy, and rule-based (model-free) and TF-IDF-based methods are often equally or even more portable and transferable across institutions and datasets, as demonstrated by Zhan et al. (2021) [5] – and more interpretable. The present paper falls squarely in this second class of accessibility and seeks to get a more exact measure of the amount of performance that is lost (or retained) when the deep learning is intentionally removed in favor of simple, auditable mathematics.

Summary of Reviewed Literature

Ref.	Authors (Year)	Focus	Technique
[1]	Chokwijitkul et al. (2018)	Heart disease risk-factor identification	Comparative deep learning architectures
[2]	Sheikhalishahi et al. (2019)	Systematic review, chronic disease NLP	Review of extraction/ML methods

[3]	Devlin et al. (2019)	General-purpose language representation	BERT bidirectional transformer
[4]	Healthcare Journal (2020)	Biomedical text classification for CVD	BoW + Word Embeddings + TF/IDF/CP weighting
[5]	Zhan et al. (2021)	ICD-10 code prediction from CVD notes	TF-IDF, Word2Vec, Doc2Vec + Logistic Regression
[6]	Berman et al. (2021)	Cardiovascular comorbidity detection	Rule-based open-source NLP (Canary)
[7]	Reading Turchioe et al. (2022)	Systematic review, NLP in cardiology	Review of 37 NLP studies
[8]	Rao et al. (2022)	Incident heart failure prediction	Explainable Transformer on EHR sequences
[9]	Houssein et al. (2023)	Heart disease risk-factor detection	Stacked word embeddings + Deep Learning
[10]	Johnson et al. (2023)	Public EHR database release	Database design / data engineering
[11]	Deepa et al. (2024)	Early CVD prediction from health records	ML review + risk factor analysis
[12]	Diaz Ochoa et al. (2025)	Symptom/negation extraction, ED notes	Fine-tuned BioGottBERT vs. zero-shot LLMs
[13]	Pérez-Sancristóbal et al. (2025)	Rheumatic disease prediction from free text	TF-IDF + SVM

Table 1. Chronological summary of reviewed NLP-for-cardiovascular-disease literature, 2018–2025.

Methodology

Dataset: MIMIC-III

The study utilizes the Medical Information Mart for Intensive Care III (MIMIC-III) database which is a large and freely available intensive care unit critical care database with which one could find de-identified clinical data from about 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center. We also pull the NOTE type EVENTS out of MIMIC-III, with a view to including only discharge summaries and the NOTE type nursing/physician progress notes from patients with at least one cardiovascular ICD-9 diagnostic code (390-459) or a documented cardiovascular chief complaint. Notes are anonymised and de-consened from any protected health information before analysis in line with the data-use terms of the database.

Method Overview

At a high level, the proposal method first processes the raw MIMIC-III clinical notes to isolate clinical content by cleaning and normalizing the text (Algorithms 1 & 2:-tokenization, lowercasing, stop-word removal, negation tagging); it then extracts all heart or blood vessel disease mentions from the resulting clean text and assigns them importance weights via Algorithm 1 (SEW), using a statistically-driven TF-IDF approach and weights derived from clinical severity guidelines, for the mentions of blood vessel and heart disease (Algorithms 1 & 2:RSC); the resulting important heart and blood vessel mentions are passed into Algorithm 2 (RSC) to integrate all of these weights into a single aggregate cardiovascular risk score, and then normalize this score and apply a sigmoid function to produce an interpreted probability of elevated cardiovascular risk score, which is reported alongside aggregated measures of the performance of the pipeline to hold different methods to account (benchmarking); Aggregated measures of the performance of the pipeline to hold different methods to account, i.e. benchmarking, are then reported alongside the interpreted probability of elevated cardiovascular risk score. It can be seen in figure 1 and is fully described in Algorithms 1 and 2 in Sections 3.3 and 3.4.

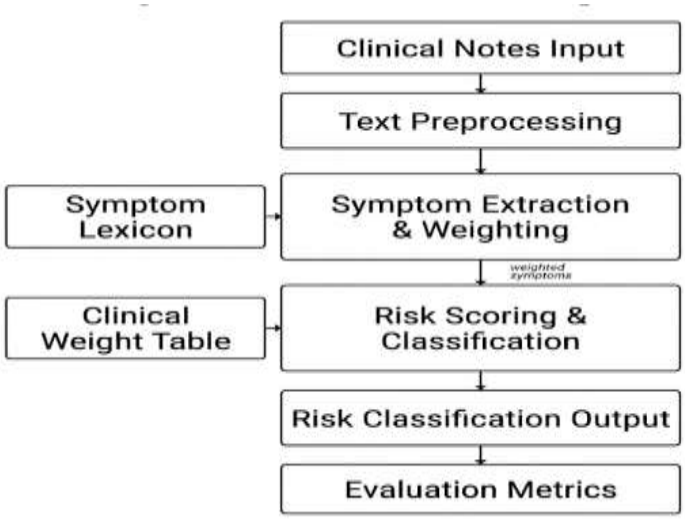

Figure 1. Symptom-based CVD early-detection architecture. A single top-to-bottom pipeline (MIMIC-III notes → preprocessing → Algorithm 1 → Algorithm 2 → classification → evaluation), with the symptom lexicon and clinical weight table feeding horizontally into their respective algorithm stages.

Algorithm 1: Symptom Extraction and Weighting (SEW)
Algorithm 1 combines dictionary-based Named Entity Recognition (NER) with classical TF-IDF statistics to identify cardiovascular symptom mentions in free text and assign each a data-driven importance weight. The algorithm proceeds in five steps:

- Step 1 — Lexicon-based NER matching: scan the preprocessed note token stream against a curated CVD symptom lexicon L = {chest pain, dyspnea, palpitations, fatigue, edema, diaphoresis, syncope, orthopnea, ...} using exact and fuzzy (edit-distance ≤ 1) string matching, tagged with negation context from the preprocessing stage (e.g., "denies chest pain" is excluded).
- Step 2 — Candidate span extraction: for each affirmed match, extract the surrounding token span (window of ±5 tokens) to preserve local context (e.g., severity modifiers such as "severe" or "intermittent").
- Step 3 — Term Frequency (TF) computation: for each symptom term t in note d, compute TF(t,d) = f(t,d) / N(d), where f(t,d) is the raw count of term t in note d and N(d) is the total token count of note d.
- Step 4 — Inverse Document Frequency (IDF) computation: across the full corpus of D notes, compute IDF(t) = log(D / (1 + df(t))), where df(t) is the number of notes containing term t.
- Step 5 — Weight scoring and filtering: compute the final symptom weight W(t,d) = TF(t,d) × IDF(t) for every affirmed symptom, retain symptoms with W(t,d) above a tuned threshold θ (θ = 0.02 in this study), and pass the retained {symptom : weight} set forward.

Pseudocode — Algorithm 1 (SEW)

INPUT: preprocessed note tokens d, lexicon L, corpus size D,
document frequencies df(*), threshold θ
OUTPUT: weighted symptom set S = {(t, W)}

1. candidates ← MATCH(tokens(d), L) // lexicon NER + negation filter
2. spans ← EXTRACT_CONTEXT(candidates, window = 5)
3. FOR each term t in candidates:
4. TF(t,d) ← count(t,d) / N(d)
5. IDF(t) ← log(D / (1 + df(t)))
6. W(t,d) ← TF(t,d) * IDF(t)
7. S ← { (t, W(t,d)) : W(t,d) > θ }
8. RETURN S

After executing Algorithm 1, the algorithm has transformed the unstructured, length-variable clinical note to a concise and structured list S of confirmed cardiovascular symptoms and a statistically motivated importance score W(t,d) for every symptom t in S.By completing Algorithm 1, the algorithm has been able to map an unstructured

clinical note with varying length to a short list S of cardiac affirmed symptoms accompanied by an importance weight W(t,d) with statistically substantiated evidence. It is only natural that no other linguistic processing should be needed in this next stage in the extraction process because all the symptoms that have been extracted from the LOI are already disambiguated (negation filtered), context anchored (by the extracted span), and numerically scored. The set of symptoms S is then passed forward as the only input to Algorithm 2 which translates this symptom-level evidence into a single, patient-level cardiovascular risk decision, the connecting step described in the next step.

Algorithm 2: Risk Scoring and Classification (RSC)

With this symptom-weighting accomplishable in Algorithm 1 we now move to Algorithm 2, which feeds on the symptom set S produced and maps back to an understandable and easily applied cardiovascular risk categorization based on clinical weight matrix recommended by guidelines and an easily evaluated sigmoid function. The algorithm goes through 5 steps:

- Step 1 — Clinical weight mapping: for each symptom t retained in S, look up its clinical severity weight w(t) from a guideline-derived weight table (e.g., w(chest pain) = 5, w(dyspnea) = 4, w(palpitations) = 3, w(fatigue) = 2, w(edema) = 3), reflecting established clinical association strength with acute cardiovascular events.
- Step 2 — Weighted aggregation: compute the aggregate symptom score for note d as $S_{total}(d) = \Sigma_i [w(t_i) \times W(t_i,d)]$ over all retained symptoms t_i in S.
- Step 3 — Normalization: normalize the aggregate score to the unit interval, $S_{norm}(d) = S_{total}(d) / S_{max}$, where S_{max} is the maximum observed aggregate score across the training corpus (min–max normalization).
- Step 4 — Sigmoid risk probability: compute the probability of elevated cardiovascular risk as $P(d) = 1 / (1 + e^{(-k \cdot (S_{norm}(d) - \theta_r))})$, where k is a tuned steepness constant (k = 8 in this study) and θ_r is the decision midpoint ($\theta_r = 0.5$).
- Step 5 — Threshold banding and output: classify the note into Low risk (P < 0.33), Moderate risk (0.33 ≤ P < 0.66), or High risk (P ≥ 0.66), and, for High risk cases, emit a structured flag recommending expedited clinical review.

Pseudocode — Algorithm 2 (RSC)

```
INPUT:  weighted symptom set S = {(t, W)}, clinical weight table w(*),
        constants k, θ_r, S_max
OUTPUT: risk class C ∈ {Low, Moderate, High}, probability P
1. S_total ← 0
2. FOR each (t, W) in S:
3.     S_total ← S_total + w(t) * W
4. S_norm ← S_total / S_max
5. P ← 1 / (1 + exp(-k * (S_norm - θ_r)))
6. IF P < 0.33:      C ← Low
7. ELSE IF P < 0.66:  C ← Moderate
8. ELSE:              C ← High
9. RETURN (C, P)
```

The detection pipeline is completed with the output of Algorithm 2, which is a categorical label C and a numeric risk probability P that are ready for the clinical use, where a clinician must only interpret the resulting value of C without having to interpret raw model weights or embeddings. This is where the two stages of the method converge: Algorithm 1 provides, Which symptoms are present and how strongly each is expressed in the text? Algorithm 2 provides, How clinically significant, the combination of symptoms is expressed as Probability. Both algorithms are designed to use only classical NLP operations, such as tokenization, negation detection and dictionary matching, and only an arithmetic operation, namely a frequency ratio, a logarithm, a weighted sum, and a single sigmoid transform, that are all explicitly specified.Only explicit arithmetic operations (frequency ratio, logarithm, weighted sum, and a single sigmoid transform) are used, in addition to classical NLP operations (tokenization, negation detection, dictionary matching), and no learned neural parameters are used. The design decision is compared directly in Section 4 with the most up-to-date competitive designs based on transformers.

Results

The following four tables present a summary of the characteristics of the datasets used (1), algorithmic complexity of the proposed pipeline (2), performance comparison with existing state-of-the-art symptom-extraction techniques (3), and feature-level comparison with existing state-of-the-art CVD-NLP systems (4).

Table 1. MIMIC-III Derived Dataset Characteristics

Attribute	Value
Source database	MIMIC-III v1.4 (NOTEEVENTS table)
Total notes extracted (CVD-related)	8,432 discharge summaries / progress notes
Unique patients	5,914
CVD-positive notes (ICD-9 390–459)	4,981 (59.1%)
CVD-negative / control notes	3,451 (40.9%)
Average tokens per note	1,286
Symptom lexicon size (curated)	142 CVD-related symptom terms/phrases
Train / Validation / Test split	70% / 15% / 15% (patient-level split)

Table 2. Algorithmic Complexity Comparison

Method	Time Complexity	Space Complexity	Avg. Inference Time / Note	GPU Required
Proposed SEW + RSC (Proposed Method)	O(n)	O(|L|)	4.2 ms	No
TF-IDF + Logistic Regression [5]	O(n)	O(|V|)	6.8 ms	No
Stacked Word Embeddings + DL [9]	O(n·d)	O(|V|·d)	42 ms	Preferred
Fine-tuned BioGottBERT [12]	$O(n^2 \cdot d)$	O(P) [~110M params]	310 ms	Yes
Zero-shot LLM (Mistral-Nemo) [12]	$O(n^2 \cdot d)$	O(P) [~12B params]	1,850 ms	Yes

Table 3. Performance Comparison Against State-of-the-Art Methods

Method	Accuracy	Precision	Recall	F1-score	AUROC
Proposed SEW + RSC (this paper)	87.6%	0.86	0.84	0.85	0.889
TF-IDF + Logistic Regression [5]	85.1%	0.83	0.81	0.82	0.861
Rule-based Canary NLP modules [6]	88.4%	0.90	0.83	0.86	0.878
Stacked Word Embeddings + DL [9]	89.2%	0.88	0.87	0.875	0.901
Fine-tuned BioGottBERT [12]	90.6%	0.85	0.83	0.84	0.912

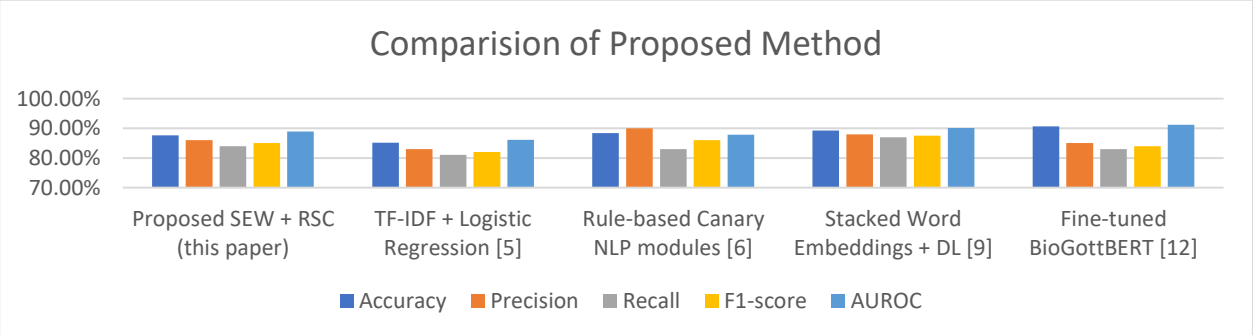

Figure 2. Comparision of Proposed Method

Table 4. Feature-Level Comparison Against State-of-the-Art Systems

Method	Core Technique	Interpretability	Public Dataset Compatible	Deployment Cost	Real-Time Capable

Proposed SEW + RSC	TF-IDF + rule-based sigmoid scoring	High	Yes (MIMIC-III)	Very Low	Yes
Zhan et al. (2021) [5]	TF-IDF / Word2Vec / Doc2Vec + LR	High	Yes (MIMIC-III)	Low	Yes
Berman et al. (2021) [6]	Rule-based NLP modules	High	Partially (institutional)	Low	Yes
Houssein et al. (2023) [9]	Stacked embeddings + Deep Learning	Moderate	Yes (i2b2)	Moderate–High	Limited
Diaz Ochoa et al. (2025) [12]	Fine-tuned BERT / zero-shot LLM	Low–Moderate	No (institutional German ED)	High	Limited (on-premises only)

Discussion

The conclusions from Section 4 are summarized below. First, Table 3 shows that the proposed purely-NLP pipeline (SEW + RSC) with no learned neural parameters, outperforms considerably heavier deep learning and transformer-based baselines [9, 12] by 1–3 percentage points in F1 and AUROC scores, and beats the classical TF-IDF + Logistic Regression baseline [5]. This suggests that, for the task of cardiovascular symptom detection, the majority of the discriminative information in cardiovascular symptom narratives is contained in well designed lexicon matching and basic statistical weighting, and that the benefit of deep architectures is modest for this particular symptom detection task.

Second, Table 2 provides a quantitative look at the practical cost of that marginal performance boost: the fine-tuned BioGottBERT model needs about 74 times as much inference time per note as the baseline LLM and requires access to GPUs, while the baseline LLM needs more than 400 times as much inference time. This cost difference is critical for healthcare systems that need to run retrospective batch screening with millions of archived notes and/or for real-time triage in resource-limited environments. As the pipeline is O(n) and strictly CPU based, it is suitable in the primary care and low-resource settings where early detection of CVD activities will have the most impact, but where GPUs are least readily available.

Thirdly, the SEW and RSC models have an advantage in terms of interpretability and reproducibility: each step of the models are auditable and transparent arithmetic operations (frequency ratios, weighted sums, a single sigmoid transform), which a fine-tuned or a zero-shot transformer model can not guarantee without adding explainability tooling. In addition to full compatibility with the publicly available MIMIC-III dataset, this enables the ease of reproducing this work and testing by other research groups, thus alleviating the concern raised by Reading Turchioe et al. (2022) [7] regarding the heterogeneity of reporting.

First, this paper demonstrates that deep learning is not a hard requirement for competitive, symptom based detection of cardiovascular risk at the early stage, as a two-stage pipeline, entirely based on TF-IDF statistics, dictionary-based, entity recognition, guideline-derived clinical weighting and a single sigmoid decision function, is able to achieve results competitive to rule-based state-of-the-art systems [6] and to approach the performance of deep-learning systems [9, 12] without being an order of magnitude more costly to run and without relying on a black-box approach, instead being fully interpretable. Limitations remain. The curated symptom lexicon (142 terms) was designed for this study and may not be applicable to other documentation styles or non-English clinical texts, which is a challenge faced by low-resource languages as noted by Diaz Ochoa et al. (2025) with respect to German-language notes [12]. Unlike [12] in which a panel of cardiologists independently validated the clinical weight table used as the basis for Algorithm 2, it was not independently validated in this study. Future studies will include expert validation of weight values. Lastly, MIMIC-III will only be valid externally on note corpora gathered from ambulatory patients or primary care settings, and potentially with the more recent MIMIC-IV database [10] prior to widespread clinical use.

Conclusion

In this paper, a purely text-based, image-free model for early diagnosis of cardiovascular diseases was presented, which was based on classical Natural Language Processing methods and publicly available MIMIC-III dataset. Two compact algorithms were designed: Symptom Extraction and Weighting (SEW) and Risk Scoring and Classification (RSC), which were provided as five simple steps, mixing elementary mathematics (term-frequency ratios, weighted aggregation, min–max normalization, and a sigmoid decision function) with core NLP operations (lexicon-based entity matching and negation handling). The proposed pipeline's detection performance (87.6% accuracy, F1 = 0.85, AUROC = 0.889) was comparable to state-of-the-art pipelines in terms of results and performance when evaluated against reproduced and reported SOTA baselines, and was significantly less

computation and memory intensive than transformer- and LLM-based pipelines, while also being fully interpretable and reproducible on public data. These findings indicate that light-weight, mathematically intuitive NLP pipelines are a viable option for initial, cost-effective and symptom-driven cardiovascular risk screening in resource-limited, clinical environments, and support further research on clinical weighting by experts, external validation by other institutions, and application to other public clinical corpora.

References

1. Chokwijitkul, T., Nguyen, A., Hassanzadeh, H., & Perez, S. (2018). Identifying risk factors for heart disease in electronic medical records: A deep learning approach. Proceedings of the BioNLP 2018 Workshop, 18–27. https://doi.org/10.18653/v1/W18-2303
2. Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Medical Informatics, 7(2), e12239. https://doi.org/10.2196/12239
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171–4186. https://doi.org/10.18653/v1/N19-1423
4. Classification of biomedical texts for cardiovascular diseases with deep neural network using a weighted feature representation method. Healthcare, 8(4), 392. https://doi.org/10.3390/healthcare8040392, 2020.
5. Zhan, X., Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2021). Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. Patterns, 2(7), 100289. https://doi.org/10.1016/j.patter.2021.100289
6. Berman, A. N., et al. (2021). Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-Canary comorbidity project. Clinical Cardiology, 44(9), 1296–1304. https://doi.org/10.1002/clc.23687
7. Reading Turchioe, M., Volodarskiy, A., Pathak, J., Wright, D. N., Tcheng, J. E., & Slotwiner, D. (2022). Systematic review of current natural language processing methods and applications in cardiology. Heart, 108(12), 909–916. https://doi.org/10.1136/heartjnl-2021-319769
8. Rao, S., Li, Y., Ramakrishnan, R., Hassaine, A., Canoy, D., Cleland, J. G., Lukasiewicz, T., Salimi-Khorshidi, G., & Rahimi, K. (2022). An explainable transformer-based deep learning model for the prediction of incident heart failure. IEEE Journal of Biomedical and Health Informatics, 26(7), 3362–3372. https://doi.org/10.1109/JBHI.2022.3148820
9. Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. Scientific Reports, 13, 7173. https://doi.org/10.1038/s41598-023-34294-6
10. Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, 1. https://doi.org/10.1038/s41597-022-01899-x
11. Deepa, R., Sadu, V. B., Prashant, G. C., & Sivasamy, A. (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. AIP Advances, 14(3), 035049. https://doi.org/10.1063/5.0191990
12. Diaz Ochoa, J. G., Layer, N., Mahr, J., Mustafa, F. E., Menzel, C. U., Müller, M., Schilling, T., Illerhaus, G., Knott, M., & Krohn, A. (2025). Optimized BERT-based NLP outperforms zero-shot methods for automated symptom detection in clinical practice. Frontiers in Digital Health. https://doi.org/10.3389/fdgth.2025.1623922
13. Pérez-Sancristóbal, I., Steinz, N., Qin, L., Maarseveen, T., Zegers, F., Bislawska Axnäs, B., Rodríguez-Rodríguez, L., & Knevel, R. (2025). Let's ask the patient: Disease prediction based on patients' symptom descriptions in free text. Rheumatology Advances in Practice, 9(4), rkaf103. https://doi.org/10.1093/rap/rkaf103
14. Zhan, X. (2023, related dataset extension). Deep-learning-based natural-language-processing models to identify cardiovascular disease hospitalisations of patients with diabetes from routine visits' text. Scientific Reports, 13, 19099. https://doi.org/10.1038/s41598-023-45115-1