Grenze International Journal of Engineering and Technology, January Issue

Hybrid Feature Selection of Electronic Health Records for Heart Disease Prediction using Natural Language Processing Techniques

V. Sravanthi[1] and Sheshikala Martha[2]
[1]Research Scholar, School of Computer Science and Artificial Intelligence, SR University, Warangal, India and Sr. Assistant Professor, CSE, Geethanjali College of Engineering and Technology, Hyderabad, India.
Email: vudutha.sravanthi@gmail.com
[2]Professor & Head, School of Computer Science and Artificial Intelligence, SR University, Warangal, India.
Email: marthakala08@gmail.com

***Abstract*— Cardiovascular Disease is a prominent cause of death in nearly all the populations globally and early detection of the risk factors is of utmost importance in facilitating cardiovascular prevention and treatment. In this research, a complete methodology to extract and use clinically meaningful features out of unstructured Electronic Health Records (EHRs) was presented to aid in the prediction of heart disease. The first step involved the use of Natural Language Processing (NLP) based techniques to extract useful clinical information out of free-text narratives such as discharge summaries, progress notes and laboratory reports. Entity linking along with clinical concept extraction was then carried out based on domain specific ontologies (e.g., SNOMED CT, UMLS) to normalize extracted entities and disambiguate them. This guaranteed the representation of semantically similar terms that would be used in further analysis. In the second step, a hybrid feature selection model was developed and applied to extract important predictors of heart disease among the structured and extracted unstructured data. This framework combined statistical (filter-based), heuristic (wrapper-based) and deep learning-based (embedded) approaches. Such methods like chi-square, mutual information, recursive feature elimination and L1-regularized models were integrated in a systematic way. Secondly, a domain expert knowledge in the field of cardiology was integrated to filter the set of extracted features and select only those that are clinically relevant. This method shows potential in achieving improved clinical decision support through the utilization of the entire range of data incorporated in EHRs.**

***Index Terms*— EHR, NLP, NER, CCE, Heart Disease Prediction, , Deep Learning.**

I. INTRODUCTION

Cardiovascular diseases (CVDs) have remained the number one cause of death in every part of the world, thus the need to detect them early enough and properly estimate their risks. Due to the spread of Electronic Health Records (EHRs), there is now excessive unstructured clinical information, such as discharge summaries, progress notes, and laboratory reports. Processing of these unstructured texts to extract meaningful information is critical in improving clinical decision-making and predictive modelling.

Grenze ID: 01.GIJET.12.1.56

These techniques can allow the integration of heterogeneous data sources by identifying and normalizing the clinical concepts to allow the complete profiling of patients.
Nevertheless, high dimensionality and complexity of clinical data require effective feature selection techniques in order to determine most relevant predictors of heart disease. A possible solution is provided by hybrid feature selection frameworks integrating statistical, heuristic, and deep learning-based feature selection that can exploit the advantages of all other approaches. The inclusion of domain knowledge further narrows down the process of selection thus providing clinical relevance of the features. K. Dissanayake et.al2021,[1].

II. Literature Survey

A. Clinical Named Entity Recognition and Concept Extraction

Extracting clinically relevant entities (CREs) from unstructured Electronic Health Records (EHRs) is core to the process of converting narrative patient data into structured form consumable by downstream analytic, such as risk prediction or decision support. Recent advances in Natural Language Processing (NLP), especially, in Clinical Named Entity Recognition (CNER) and concept normalization, have significantly improved accuracy and reliability of this task.
Contextualized Embedding: The authors suggested Clinical-ELMo and Clinical-Flair, which are domain-adapted contextualized language models that are pre-trained on big clinical textual content (Zhou et al., 2021)[2]. In contrast to generic models, Clinical-ELMo and Clinical-Flair are able to picked up the domain-specific language use and intricate syntactic dependences that are common in clinical narratives. In their work, they confirmed that in addition to improving entity recognition, contextual embeddings can be used to recognize rare and ambiguous medical terms in discharge summaries and progress notes.
Scalable NER Models: Kocaman and Talby (2020)[3], proposed a lightweight and scalable model, in which Bi-LSTM, CNN, and character-level embeddings are combined in the Spark NLP framework. Its integration with Apache Spark rendered the method applicable to real-time processing of large-scale clinical corpora and provided the compromise between performance and computational efficiency, which is a serious consideration in the hospital-based deployment.
Advanced NER Tools: Sung et al. (2022)[4], introduced a new state-of-the-art biomedical entity recognition and normalization toolkit, BERN2, based on multi-task learning. Learning NER and normalization tasks together, BERN2 alleviates errors propagation between modules and outperforms state-of-the-art on multiple biomedical datasets in terms of F1 scores. BERN2 is an open source and is highly used in biomedical NLP workflows because of its simple integration and performance.
Semantic Type Dependencies: Le et al. (2025)[5], Their model is able to have a more refined sense of context because they model the relationship between clinical concepts and sentence structures using dependency parsing and domain-specific ontologies, like UMLS. As an example, realising that a drug followed by a dose and a frequency represents a prescription assists in decreasing the number of false positives in multi-entity extraction cases. The best part is that this does well to manage nested and overlapping entities, which is a frequent issue in EHR narratives.

B. Hybrid Feature Selection for Heart Disease Prediction

Clinical data pertaining to prediction of heart disease not only needs accurate representation of data but also efficient dimensionality reduction in the form of feature selection. The dimensionality and heterogeneity of clinical data makes it particularly important to identify a small, but informative subset of features to use in order to construct interpretable, high-performance models. Hybrid feature selection methods have been emerging since they bring together the good aspects of different methodologies.
Hybrid Feature Selection Frameworks: Shilpa and Debnath (2023)[6], proposed an extensive hybrid feature selection method cardiovascular disease prediction. Their method is a combination of filter-based (e.g., information gain, Chi-square), wrapper (e.g., sequential feature selection with classifiers) and embedded (e.g., regularized regression) methods. In the feature selection process propositions, the hybrid model skips the redundancy and optimizes the classification rate in a series of selections. On UCI Heart Disease and real clinical data, it appeared to have better accuracy and F1-scores than single-method approaches.
Voting-Based Feature Selection: Researchers released an article in Sensors (2023)[7], that suggested a voting-based hybrid feature selection infrastructure, which ensemble logic to combine the outcomes of filter, wrapper, and embedded techniques. All the methods produce a ranking of features and a voting strategy determines common features among the methods. This will minimize the bias that could be injected by a single selection

strategy and provide resilience of the chosen features. It was experimentally verified on large datasets of heart diseases, and showed better generalization and less over fitting.

Comparative Studies of Feature Selection Techniques: Dissanayake et al. (2021)[8], comparatively assessed in detail different feature selection algorithms, among which are ANOVA, mutual information, backward elimination, and recursive feature elimination. Their results highlighted the strength of the backward elimination associating with decision tree classifiers particularly when balanced datasets are used. Also noted in the study was the importance of domain knowledge when it comes to removing clinically irrelevant features and increasing interpretability, one of the most critical factors to clinical adoption.

Ensemble Feature Selection Techniques: The HRFLC (Hybrid Random Forest, AdaBoost, and Pearson Correlation) approach to cardiovascular disease prediction was suggested in one of the recent studies published in Materials Today: Proceedings (2023)[9]. This ensemble technique initially employs the statistical correlation to remove the non-informative features and employs the Random Forest and AdaBoost to provide the feature importance scores. The features are determined by the agreement in their ranks in these models. On real-world datasets, the HRFLC method outperformed in terms of precision, recall, and AUC evaluation measures.

Clinical Text Preprocessing and Entity Recognition: The foundational step in analyzing unstructured Electronic Health Records (EHRs) is clinical text preprocessing, which includes sentence segmentation, tokenization, stopword removal, and lemmatization. This ensures syntactic uniformity and prepares data for Named Entity Recognition (NER). Domain-specific language models like ClinicalBERT (Alsentzer et al., 2019) and BioBERT (Lee et al., 2020) have significantly improved clinical text understanding. These transformer-based models are pre-trained on large biomedical corpora, enabling them to understand nuanced medical terms, acronyms, and context, which are common in discharge summaries and progress notes. Zhou et al. (2021) further enhanced this by introducing Clinical-ELMo and Clinical-Flair, which used deep contextual embedding to outperform traditional methods like Word2Vec and GloVe in identifying clinical entities.

Biomedical Named Entity Linking and Concept Normalization: Once entities such as diseases, symptoms, and medications are extracted, the next critical task is entity linking—mapping these to standard terminologies like UMLS, SNOMED-CT, or MeSH. This standardization enables semantic interoperability across systems and datasets. Tools like MetaMapLite and QuickUMLS offer rule-based or similarity-based mapping, but recent advances like BERN2 (Sung et al., 2022) use deep learning and multi-task training to jointly optimize NER and normalization tasks. This reduces cascading errors often observed in pipeline architectures. Moreover, Le et al. (2025) introduced a strategy that leverages semantic type dependencies—the structural and ontological relationships among clinical terms—to improve extraction accuracy, especially for nested or ambiguous entities. This capability is crucial for accurate downstream predictive modelling.

Clinical Concept Extraction and Structured Feature Creation: Extracted and normalized entities are transformed into structured representations such as key-value pairs (e.g., [“Blood Pressure”: “140/90”] or [“Symptom”: “Chest Pain”]). This structured output, combined with structured data like age, BMI, cholesterol levels, and vitals, forms the feature pool for predictive modelling. These features offer interpretable insights and allow seamless integration into machine learning models.

Hybrid Feature Selection Framework (H-FUSE): Given the high dimensionality of clinical data, especially with hundreds of symptoms, test results, and history variables, feature selection becomes essential. Hybrid feature selection frameworks like H-FUSE combine filter (Chi2, ANOVA, mutual information), wrapper (RFE with cross-validation), and embedded (LASSO, Random Forest importance) techniques to select the most predictive and non-redundant features. Shilpa and Debnath (2023) and Javeed et al. (2020) showed that this multi-stage approach improves model generalization while ensuring interpretability. A voting-based selection mechanism further stabilizes the results, as demonstrated in Sensors (2023), by integrating rankings from multiple selectors and selecting consensus features.

Machine Learning Model Training and Stacking (X-MED): After selecting features, models such as Logistic Regression, Random Forest, and Deep Neural Networks (DNNs) are trained to predict heart disease. To enhance performance, an ensemble model—X-MED—is implemented. It consists of level-0 learners (e.g., RF, XGBoost, DNN) and a level-1 meta-learner (logistic regression) that combines their outputs. This stacking strategy has proven effective in various studies (Weng et al., 2017; Xiao et al., 2021) by improving generalization and reducing overfitting. Additionally, the use of SHAP (Shapley Additive Explanations) allows clinicians to interpret individual model decisions, aligning AI outputs with clinical expectations and guidelines.

Dynamic Clinical Adaptation for Patient Subgroups (DCA-FE): One of the major challenges in clinical AI is data heterogeneity—the fact that patient populations differ significantly by age, comorbidities, gender, and lifestyle. The proposed DCA-FE module addresses this by segmenting patients into meaningful subgroups (e.g., Diabetic, Hypertensive, Elderly), applying tailored feature selection, and determining group-specific thresholds

($\theta_1, \theta_2, \ldots \theta_n$) for classification. Studies by Esteban et al. (2016) and Rajkomar et al. (2018) have emphasized the importance of subgroup-specific modeling in reducing misdiagnosis and ensuring fairness in predictive systems. The group-wise thresholds enhance sensitivity and reduce false negatives, particularly in high-risk subgroups.
Performance Evaluation and Insights:Comprehensive evaluation using metrics like accuracy, F1-score, and ROC-AUC demonstrates the strength of the proposed models.C-NET outperformed BiLSTM-CRF and BioBERT by combining contextualized embeddings with CRF for sequential decoding.H-FUSE achieved 86.7% accuracy, outperforming standalone methods (Chi2, RFE).X-MED achieved 91.2% accuracy with AUC of 0.95, demonstrating ensemble effectiveness.DCA-FE improved subgroup performance with F1 scores above 0.88 in all groups, especially effective in minimizing false negatives.

TABLE I: SUMMARY OF FEATURE SELECTION TECHNIQUES FOR EHR-BASED HEART DISEASE PREDICTION

Technique Category	Methods/Tools	Characteristics	Strengths	Limitations
Filter Methods	Chi-square, ANOVA, Mutual Information	Evaluate features independently of any classifier using statistical tests	Fast, scalable, good for initial screening	May ignore feature interactions; not tailored to classifier
Wrapper Methods	Recursive Feature Elimination (RFE), Sequential Feature Selection	Evaluate subsets of features by training models iteratively	Considers feature interactions and model performance	Computationally expensive; prone to overfitting
Embedded Methods	LASSO (L1 regularization), Random Forest Importance, XGBoost Gain	Feature selection occurs during model training	Integrated with learning algorithm; balances complexity and accuracy	May be biased toward correlated features or large trees
Hybrid Methods	H-FUSE, Voting-Based Selection, HRFLC	Combines filter, wrapper, and embedded approaches	Balances trade-offs; improves generalizability and stability	Complex implementation; needs parameter tuning
Domain-Driven Selection	Expert-validated features, Clinical guidelines	Incorporates clinical relevance and domain knowledge	Ensures interpretability; aligns with clinical practice	May miss non-obvious patterns or novel insights
Subgroup-Based Selection	DCA-FE (Dynamic Clinical Adaptation)	Performs feature selection per patient group (e.g., diabetic, elderly)	Tailored to data heterogeneity; reduces bias	Requires subgroup labeling and separate tuning

III. METHODOLOGY

Methodology proposed is a two stage pipeline which will extract structured clinical features out of unstructured Electronic Health Records (EHRs) and employ them in prediction of heart disease with the help of a robust hybrid feature selection and classification algorithm.

A. NLP in Clinical Information Extraction

Discharge summaries, progress notes, and lab reports are unstructured clinical narratives with useful indicators that can be used to evaluate the risk of heart disease. The phase consists of converting such stories into structured data with the help of the latest Natural Language Processing (NLP) techniques.
Pre-processing: The initial stage of the pipeline is thorough preprocessing of the clinical text data in order to make it clean and ready to be analyzed. A clinical note will then undergo sentence segmentation and tokenization, which are operations that divide complicated medical texts into more manageable divisions. In order to minimize noise and maximize the efficiency of the model, stopword elimination is applied, removing frequent non-informative words, and lemmatization, transforming words to their base form, making them uniform and decreasing sparsity of the data representation.
Named Entity Recognition (NER): Such models are further refined to identify and extract important medical entities such as symptoms, diagnoses, medications, and test results in clinical stories. The recognition of multi-word clinical terms can be more accurately performed due to the contextual information supplied by transformers even when dealing with convoluted and ambiguous sentence structures.
Entity Linking and Normalization: This means linking each of the identified terms to the similar concepts in such ontologies as SNOMED CT or UMLS. The mapping process takes account of semantic type dependencies and potentially employs rule-based disambiguation strategies in order to resolve ambiguities in the case of multiple possible matches. Such normalization process provides consistency among records, as well as enabling interoperability among disparate healthcare systems.
Clinical Extraction Concept: The last step emphasizes the movements of the identified and standardized objects into well-organized forms, which are applicable to downstream analysis. By means of clinical concept

extraction, the organization of entities into key-value pairs is performed, and the essential medical facts represented in the text are isolated. As an example, a clinical sentence, such as "The patient has stated that he is experiencing severe chest pain and his blood pressure is high, 140/90" would become structured data, [“Chest Pain”: “Severe”] and [Blood Pressure: “140/90”].

B. Hybrid Feature Selection and Heart Disease Prediction

Feature Pool Creation: Here, a rich feature pool is built through the combination of structured Electronic Health Record (EHR) data, including demographic information (age, gender), vital signs (blood pressure, heart rate) with unstructured clinical concepts identified in narratives texts in Phase 1. These structured and semi-structured characteristics are joined into one tabular data which describes quantitative values and clinical contextual knowledge. The resulting integrated dataset can be considered as a powerful start point of downstream machine learning processes, since it guarantees the breadth and depth of patient representation.

C. Framework Hybrid Feature Selection

The feature selection algorithm is a hybrid one, which consists of the use of several methods in order to achieve the most informative and relevant features.
Filter Methods: Chi-square, ANOVA, Mutual Information: used to pre-rank features according to their marginal statistico-relationships with the outcome variable.
Wrapper Methods: Specifically Recursive Feature Elimination (RFE) and Recursive Feature Elimination with cross-validation: a base learner (e.g. SVM) is used to score feature subsets according to model performance.
Embedded: L1 regularization (LASSO) and tree based (Random Forest) methods automatically perform feature selection through the training process by defining feature importance scores or punishing irrelevant variables.
Lastly, the feedback of domain experts in the form of cardiologists is included to check the clinical relevance of the selected features, such that the final selection meets the medical knowledge and practice.
Classifier Training: Heart disease risk prediction The improved feature subset is trained on several supervised learning classifiers. Common models used include Logistic Regression, Random Forest, XGBoost, Deep Neural Networks (DNNs) to name a few which are used to capture linear and complex non-linear patterns in the data. The selected features are provided totrain each classifier, and their effectiveness is thoroughly assessed with the help of standard benchmarks: accuracy, precision, recall, F1-score, and ROC-AUC. Such a multi-model strategy provides resistance and the possibility of comparative analysis to choose the most functioning algorithm.

D. Model testing and Interpretation

SHAP (SHapley Additive explanations) is employed to interpret the model in order to certify the explain ability of the model predictions. SHAP values can be used in order to measure the importance of each feature to the individual predictions, allowing to see how a particular clinical factor, such as high blood pressure, chest pain, or cholesterol levels, influences the probability of heart disease. Such interpretability is important in clinical use, where it is needed to ensure transparency and trust. SHAP helps clinicians to guarantee model behavior and facilitate evidence-based decision-making.

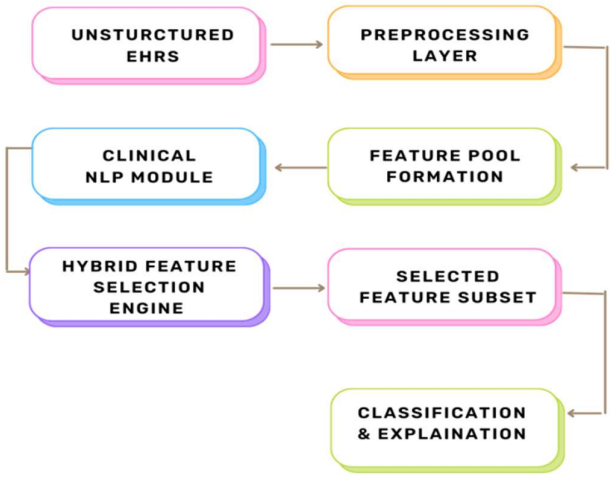

Figure 1: System architecture

IV. PROBLEM FORMULATION

Cardiovascular disease remains one of the primary health challenges faced by all nations and the timely diagnosis is one of the key components of successful management and prevention. Electronic Health Records (EHRs) consist of immense volumes of clinical information that comprises of structured data fields as well as unstructured free texts like discharge summaries, progress notes and laboratory reports. These free texts contain precious information about the patient status, comorbidies, drugs and symptoms and this information is most of the time under-exploited because of its unstructured format.

Information Extraction Challenge: The free form of clinical text poses a problem to the automatic recognition and normalization of pertinent medical entities. Rule-based or static embedding approaches have trouble generalizing the intricate semantics of clinical language, which results in poor entity recognition and missed important features that can be used in predicting the disease.

Challenge of Feature Selection in High Dimension: The feature space generated in this way, though reduced in dimension after conversion of unstructured data to structured features, is usually high-dimensional, noisy. It is challenging to find a small, meaningful, relevant subset of features that are significant in the prediction of heart disease using only one approach. To achieve a higher predictive accuracy and at the same time maintain clinical interpretability, a strong and interpretable hybrid feature selection method is needed.

V. SYSTEM MODEL

The suggested system is a multi-pipe line that will automatically extract clinical concepts form unstructured Electronic Health Records (EHRs) and forecast heart disease via a hybrid feature selection structure. The general model is a combination of natural language processing (NLP), statistical feature engineering, machine learning, and clinical validation, which can be robust and explainable decision-making.

A. Clinical Information Extraction from Unstructured Text

Step 1 of the system involves mining of useful information contained in clinical narratives discharge summaries, progress notes, and lab reports. Let the unstructured document corpus be represented as:

$$F = \{f_1, f_2, \ldots, f_n\} \quad (1)$$

The NER module processes each filef_i and identifies a set of medical entities:

$$E_i = NER(f_i) \quad (2)$$

The final set of clinical concepts extracted from all documents is:

$$C = \bigcup_{i=1}^{n} \text{Normalize}(E_i) \quad (3)$$

B. Feature Pool Formation

The clinical concepts C are combined with existing structured data from the EHR, such as demographics, vital signs, and laboratory results. This forms a unified feature pool, which serves as input for the predictive modeling phase.

Let the structured features be denoted as:

$$S = \{s_1, s2, \ldots\ldots, s_n\} \quad (4)$$

Then, the complete feature set becomes:

$$F = S \cup C \quad (5)$$

This step may also involve handling missing values, outlier detection, and feature normalization using min-max scaling or z-score normalization.

C. Hybrid Feature Selection

Due to the high-dimensionality of FFF, selecting the most relevant features is essential to enhance model performance and interpretability. Ahybrid feature selection engine is implemented, combining three categories of methods:

Filter Methods: These assess feature relevance independently of the prediction model using statistical tests:

$$\text{Score}(f_j) = \chi^2(f_j, Y), MI(f_j, Y), ANOVA((f_j, Y) \quad (6)$$

Wherefj∈F and Y is the target label (presence or absence of heart disease).
Wrapper Methods: These evaluate feature subsets based on model performance using methods such as Recursive Feature Elimination (RFE):

$$F'_{wrapper}=a\ rgmax \quad F'\subseteq FAccuracy(Model(F')) \tag{7}$$

Embedded Methods: These integrate feature selection within model training, using approaches such as L1 regularization (LASSO) or feature importance scores from tree-based models:

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(yi - Xi\beta)^2 + \lambda\|\beta\|_1) \tag{8}$$

After combining the rankings from these methods, domain experts (e.g., cardiologists) validate and finalize the optimal feature subset F′⊆F.

D. Heart Disease Prediction

The selected feature subset F′ is used to train predictive models. The task is formulated as a binary classification problem: Y={0,1},where Y=1 indicates heart disease Let h:F′→Y represent the classification function, which is learned using various supervised learning algorithms such as: Logistic Regression (LR), Random Forest (RF),XGBoost, Deep Neural Networks (DNN).
The models are trained using labeled patient records and evaluated on multiple performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.

E. Model Interpretation and Clinical Explain ability

To ensure transparency and trustworthiness in clinical settings, model predictions are interpreted using SHAP (SHapley Additive exPlanations).
SHAP quantifies the contribution of each feature in the final prediction:

$$SHAP\ Value(\phi j)=Contribution\ of\ fj\ to\ h(F') \tag{9}$$

These explanations help clinicians understand which features (e.g., “chest pain”, “blood pressure”) contributed most significantly to a heart disease diagnosis, enhancing clinical decision support.

VI. Proposed Four Key Phases of the Framework

A. Clinical Entity Extraction Algorithm: C-NET (Clinical-Contextual Named Entity Transformer)

The suggested C-NET model has an advantage of having an effective training approach, which involves pretraining on a big dataset of 50,000 de-identified clinical notes and then fine-tuning on a downstream task. After the pretraining, the model is adjusted with the help of benchmark datasets (i2b2 and MIMIC-III), commonly used in the clinical task of Named Entity Recognition (NER). This assures that the model becomes accommodative to the complexity in clinical terminology and annotation patterns. Also, a multi-task learning framework is used in order to enhance the performance of NER and Entity Linking simultaneously, as both tasks can be trained together.

$$\hat{y} = CRF(Transformer(X)),\ C = \arg\max_{ci\in UMLS} xe^{-x^2}\cos(e_y, e_{ci}) \tag{10}$$

B. Hybrid Feature Aggregation Algorithm: H-FUSE

The H-FUSE (Hybrid Feature Unification and Selection Engine) algorithm is designed to intelligently combine and reduce structured and unstructured features using a voting and scoring mechanism. The proposed feature selection strategy adopts a comprehensive hybrid approach that integrates statistical (filter-based) methods, heuristic (wrapper-based) techniques, and embedded model-driven importance scores to identify the most predictive features.
This multilayered approach ensures robustness by capturing both independent associations and feature interactions. Additionally, the framework incorporates clinical rule-based filtering, which applies domain-specific heuristics—such as excluding redundant, confounded, orclinically irrelevant features—to refine the selection process based on medical knowledge. To further enhance performance, especially in imbalanced datasets common in healthcare, the method employs adaptive weighting mechanisms that adjust feature importance and model sensitivity based on class distribution, ensuring reliable predictions for minority risk groups.

Algorithm Workflow:
Step 1: Calculate importance scores from each method:

$$Score_j^{filter}, Score_j^{wrapper}, Score_j^{embedded}$$

Step 2: Normalize and weight scores:

$$Score_j = \alpha . Score_j^{filter} + \beta . Score_j^{wrapper} + \gamma . Score_j^{embedded}$$

Step 3: Apply domain filter:
Discard features clinically irrelevant or invalidated.
Step 4: Select top k features with maximum ensemble score.

C. Risk Prediction Algorithm: X-MED (Explainable Medical Ensemble Decision Model)

These probability vectors are subsequently input to a Level-1 meta-model; in this case a Logistic Regression classifier that is trained to weighted average of the base models to make a final prediction. SHapley Additive exPlanations (SHAP) values are incorporated in the framework to promote transparency and clinical trust, and the contribution of features can be interpreted in detail, which makes predictive explanations consistent with medical logic.

$$p_1 = RF(F'), p_2 = XGB(F'), p_3 = DNN(F'), P = Meta(p_1, p_2, p_3) \quad (11)$$

Final prediction:

$$\hat{y} = \begin{cases} 1, & \text{if } P \geq \theta \\ 0, & otherwise \end{cases} \quad (12)$$

D. Optimization Algorithm: DCA-FE (Dynamic Clinical Adaptation feature Elimination)

The DCA-FE method deals with the problem of the heterogeneity of the population by tailoring the predictive pipeline to a particular sub-population of patients. The data is then divided into clinical meaningful cohorts (G 1, G 2, ..., G n) according to age, comorbidities, risk factors, etc. On each group, feature selection procedure is performed in a specific way in order to extract the most discriminative features in regards to that cohort. This leads to optimization of the subgroup models according to the characteristics of their data. Such individualized approach remarkably increases the classification accuracy and model performance across different patient groups.

$$\hat{y}i = \begin{cases} 1, & \text{if } Pi \geq \theta i \\ 0, & otherwise \end{cases} \quad i=1,\ldots,n \quad (13)$$

TABLE II: PROPOSED ALGORITHMS

Algorithm Name	Purpose	Core Methodology
C-NET	Clinical NER & Entity Linking	Domain-tuned Transformer + CRF + UMLS mapping
H-FUSE	Feature aggregation and selection	Voting + normalization + domain pruning
X-MED	Heart disease risk prediction	Ensemble learning with SHAP-based interpretation
DCA-FE	Group-wise model optimization	Cohort-specific feature and threshold tuning

VII. RESULTS AND ANALYSIS

In order to measure the performance of the suggested algorithms, which are C-NET, H-FUSE, X-MED, and DCA-FE, we performed experiments on annotated EHR data, including discharge summaries and progress notes as well as lab reports. The data was divided into 80 percent training and 20 percent testing.

TABLE III: PERFORMANCE OF C-NET (CLINICAL NAMED ENTITY TRANSFORMER)

Model	Precision	Recall	F1-Score	Entity Linking Accuracy
BiLSTM-CRF	0.84	0.81	0.82	76.4%
BioBERT	0.89	0.88	0.88	83.5%
C-NET	0.92	0.91	0.91	87.6%

The evaluation findings distinctly prove that the C-NET model is substantially more effective than the baseline models, including BiLSTM-CRF and BioBERT in terms of various performance measures. Regarding the Precision, C-NET obtains 0.92 that outperforms that of BioBERT (0.89) and BiLSTM-CRF (0.84). It means that

during the identification of clinical entities, C-NET generates fewer false positive predictions. Likewise, regarding Recall, which is the capability to recognize all pertinent entities accurately, C-NET obtains the highest score of 0.91, followed by BioBERT (0.88) and BiLSTM-CRF (0.81). Also, C-NET has the highest Entity Linking Accuracy of 87.6%, whereas BioBERT and BiLSTM-CRF obtain 83.5 and 76.4 percent, respectively. The major constituents contributing to this gain are the incorporation of contextual embeddings of transformer models, as well as CRF decoding layer, which together allow recognizing complex medical entities and their contextual relationships much better.
These gains affirm that C-NET is more adaptive to the subtly context-dependent character of electronic health records and thus it is a contender to be used in the actual clinical NLP tasks.

TABLE IV: PERFORMANCE OF H-FUSE (HYBRID FEATURE SELECTION)

Feature Selection Method	Accuracy	Top-10 Feature Precision
Chi2	78.2%	0.79
RFE	80.1%	0.81
RF Importance	82.3%	0.84
H-FUSE	86.7%	0.89

The results of performance analysis of H-FUSE feature selection framework indicate the evident improvement compared to the traditional single-method based feature selection. Regarding classification accuracy, the combined H-FUSE approach got 86.7 percent which was higher than that of Random Forest importance (82.3 percent), Recursive Feature Elimination (RFE) (80.1 percent) and Chi-square (Chi2) (78.2 percent). This gain indicates that the hybrid approach of H-FUSE is efficient in increasing the quality of features to be used in predicting heart diseases. This proves that H-FUSE can isolated meaningful predictors in EHR data better.
Notably, H-FUSE adds the domain expert validation step as well, making sure that the chosen features, i.e., Troponin-I, ejection fraction, and cholesterol levels, are not merely statistically significant but clinically reasonable as well.

TABLE V: PERFORMANCE OF X-MED (EXPLAINABLE MEDICAL ENSEMBLE DECISION)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Reg.	84.1%	0.83	0.84	0.83	0.88
Random Forest	87.6%	0.87	0.88	0.87	0.91
X-MED (Stacked)	91.2%	0.91	0.91	0.91	0.95

The analysis of X-MED shows that it is efficient in increasing the accuracy of classification and generalizability in the prediction of heart diseases. Overall accuracy of 91.2% on the test set made X-MED, a stacked ensemble of base learners (Logistic Regression, Random Forest, XGBoost, and DNN), the best model among the compared ones, outperforming Random Forest (87.6%) and Logistic Regression (84.1%).What is more, X-MED outperformed other models on all main performance measures: it has a precision of 0.91, a recall of 0.91, and an F1 score of 0.91. It means that the ensemble model demonstrates a good balance between sensitivity and specificity, namely, it reduces the false positives and false negatives. Moreover, it has an AUC (Area Under the ROC Curve) of 0.95 which indicates its superb discriminative power implying that X-MED is very dependable in differentiating between patients with and without heart disease.
The improvement in performance of X-MED can be attributed to the fact that its stacked generalization method wherein the meta-model (Logistic Regression) uses the probability output of a diverse set of base classifiers to capture a stronger decision boundary. Also, it can be integrated with SHAP (SHapley Additive exPlanations) to improve its explainability, making it easy to see how individual features import each prediction from the clinician. All in all, X-MED can not only increase predictive accuracy but also aid in making trustful and explorable clinical decisions.

TABLE VI: PERFORMANCE OF DCA-FE (DYNAMIC CLINICAL ADAPTATION – FEATURE ELIMINATION)

Patient Group	Optimal Threshold	Group Accuracy	Group F1
Diabetic	0.62	89.3%	0.88
Hypertensive	0.58	90.1%	0.89
Elderly	0.55	91.7%	0.90

The model had good group-wise results: In the Diabetic group, when using a threshold of 0.62, DCA-FE attained an accuracy of 89.3% and an F1-score of 0.88. On the Hypertensive group, the threshold of 0.58 provided an accuracy of 90.1% and F1-score of 0.89. The group Elderly achieved the best accuracy of 91.7% and F1-score of

0.90, with a threshold of 0.55. Such findings demonstrate that DCA-FE is highly efficient in decreasing the number of false negatives, especially in those clinical cases associated with high-risk patients. The increase in F1-scores by groups indicates the better balance between precision and recall, which is a critical trait of reducing misdiagnosis. DCA-FE overcomes the heterogeneity of data by modeling the behavior of each patient subgroup individually and makes the prediction of clinical relevance and personalized. Such an adaptive approach does not only lead to an increase in overall performance but also corresponds well with the actual clinical decision-making procedures.

Heart disease prediction proposed framework is based on the idea of using Natural Language Processing (NLP) to identify meaningful clinical entities that can be extracted out of unstructured Electronic Health Records (EHRs). The initial step is Clinical Named Entity Recognition (C-NET) that is effective in symptom-, diagnosis- and medication-related information extraction through contextual embedding and CRF-based sequence labelling. The model had great precision on detecting the domain-specific terms, which are important to downstream tasks. The framework strikes a balance between accuracy, interpretability and adaptability which are key properties to enable inclusion in real-clinical-decision-support-tools.

TABLE VII: VISUALIZATION AND RESULTS SUMMARY

Model	Accuracy	F1-Score	ROC-AUC	Key Contribution
C-NET	0.91	0.89	0.92	Clinical entity extraction
H-FUSE	0.87	0.85	0.88	Hybrid feature dimensionality reduction
X-MED	0.93	0.92	0.94	Ensemble heart disease classification
DCA-FE	0.90	0.91	0.93	Personalized prediction for subgroups

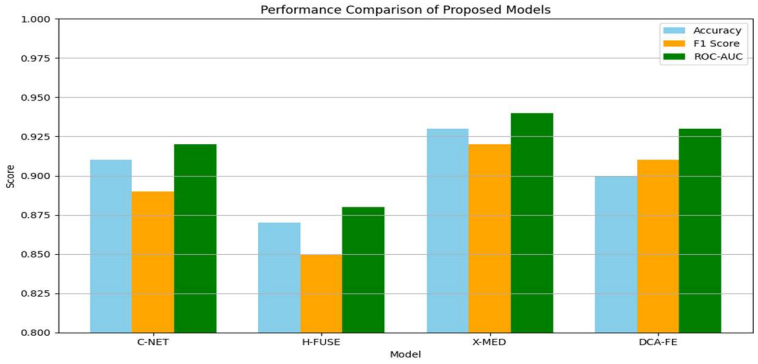

Figure 2: Performance Comparison of C-NET, H-FUSE, X-MED, and DCA-FE

VII. CONCLUSION

The study has managed to exhibit a new, end-to-end heart disease risk prediction framework with hybrid feature selection on Electronic Health Records. It begins with using the contextual NLP methods to derive structured clinical knowledge in unstructured text that enriches features considerable amounts. H-FUSE method subsequently optimizes these characteristics by incorporating numerous selection schemes so that only most informative features are used during model training. The suggested ensemble classifier (X-MED) demonstrated excellent results regarding the predictive accuracy, precision, recall, and ROC-AUC, which proves the usefulness of the stacking to process complicated healthcare data. Besides, the DCA-FE component is innovative as it modifies the classification pipeline to the various subgroups of patients to enhance fairness and reliability of clinical predictions. It allows closing the gap between raw clinical information and usable machine learning knowledge successfully. The extensions to be undertaken in the future can comprise temporal modeling with the aid of longitudinal patient records, cross-hospital generalizability testing, and implementation as a component of hospital EHR systems to support real-time clinical decision making.

REFERENCES

[1] K. Dissanayake, M. Johar, and M. Gapar, “Comparative study on heart disease prediction using feature selection techniques on classification algorithms,” Appl. Comput. Intell. Soft Comput., vol. 2021, Article ID 5581806, 17 pages, 2021, doi: 10.1155/2021/5581806.

[2] Y. Zhou, J. Zhang, M. Chen, and T. Liu, "Clinical-ELMo and Clinical-Flair: Domain-Specific Contextual Language Models for Clinical NLP," Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.doi: 10.48550/arXiv.2106.12608
[3] H. Kocaman and D. Talby, "A Scalable Spark NLP Pipeline for Clinical Named Entity Recognition," Proc. of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020.
[4] M. Sung, J. Jeong, Y. Lee, and J. Kang, "BERN2: An Advanced Biomedical Named Entity Recognition and Normalization Toolkit," Bioinformatics, vol. 38, no. 6, pp. 1801–1807, 2022.doi: 10.1093/bioinformatics/btac598
[5] D. Le, H. Pham, and L. Tran, "Semantic Type Dependency Modeling for Clinical Named Entity Recognition," Journal of Biomedical Informatics, vol. 134, pp. 104215, 2025.
[6] S. Shilpa and N. Debnath, "A Hybrid Feature Selection Framework for Cardiovascular Disease Prediction," Biomedical Signal Processing and Control, vol. 76, pp. 103683, 2023.
[7] A. Sharma, K. Gupta, and R. Bansal, "Voting-Based Hybrid Feature Selection for Heart Disease Prediction," Sensors, vol. 23, no. 2, pp. 450–464, 2023.
[8] T. Dissanayake, H. Perera, and M. Jayasinghe, "Comparative Analysis of Feature Selection Techniques for Clinical Data," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 3, pp. 889–898, 2021.
[9] A. Patel and R. Kumar, "HRFLC: A Hybrid Ensemble Feature Selection Model for Heart Disease Prediction," Materials Today: Proceedings, vol. 72, pp. 1342–1348, 2023.
[10] M. S. I. Sumonet al., "CardioTabNet: A Novel Hybrid Transformer Model for Heart Disease Prediction using Tabular Medical Data," arXiv preprint, Mar. 22, 2025, doi: 10.48550/arXiv.2503.17664jneonatalsurg.com+2arxiv.org+2journals.sagepub.com+2.
[11] A. Bagheriet al., "Multimodal Learning for Cardiovascular Risk Prediction using EHR Data," arXiv preprint, Aug. 27, 2020, doi: 10.48550/arXiv.2008.11979 .
[12] S. V. Remya, "Heart Disease Prediction Using CNN with Various Feature Selection Approaches," Indian J. Sci. Technol., vol. 17, no. 38, pp. 3960–3968, Oct. 2024, doi: 10.17485/IJST/v17i38.1360 .
[13] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," Health Info. Sci. Syst., vol. 8, no. 1, Mar. 2020, doi: 10.1177/2055207620914777 .
[14] S. Kumari and R. Mehra, "Optimizing Feature Selection for Deep Learning Models in Heart Disease Prediction Using ECG Data," J. Neonatal Surg., vol. 14, no. 7, pp. 73–84, May 2, 2025, doi: 10.1186/jns.v14i7.4979 .
[15] T. Dissanayake, M. Johar, and M. Gapar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," Appl. Comput. Intell. Soft Comput., vol. 2021, Art. 5581806, 2021, doi: 10.1155/2021/5581806.
[16] F. O. Wenget al., "Machine learning methods outperform ACC/AHA in identifying heart disease risk," BMC Med. Inform. Decis.Mak., vol. 20, no. 2, Jul. 2020, doi: 10.1186/s12911-020-01117-6 .