JOIV : Int. J. Inform. Visualization, 8(4) - December 2024 2287-2296

INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

A Model for Enhancing Pattern Recognition in Clinical Narrative Datasets through Text-Based Feature Selection and SHAP Technique

Sirajo Muhammad Dalhatu [a,*], Masrah Azrifah Azmi Murad [a]

[a] *Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia*

Corresponding author: [*]siraj9169@gmail.com

Abstract*—Clinical narratives contain crucial patient information for predicting cardiac failure. Accurate and timely cardiac failure recognition (CFR) significantly impacts patient outcomes but faces challenges like limited dataset sizes, feature space sparsity, and underutilization of vital sign data. This study addresses these issues by developing a methodology to improve CFR accuracy and interpretability within clinical narratives. Four datasets—the Framingham Heart Study, Heart Disease from Kaggle, Cleveland Heart Disease, and Heart Failure Clinical Records—undergo preprocessing, including handling missing values, removing duplicates, scaling, encoding categorical variables, and transforming unstructured data using natural language processing (NLP). Various feature selection methods (Chi-Squared, Forward Selection, L1 Regularization) are used to identify influential features for CFR, and the SHapley Additive exPlanations (SHAP) technique is integrated to improve interpretability. Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) models are trained and evaluated. Performance was evaluated using accuracy, precision, recall, f1-score, and area under the receiver operating characteristic curve (AUC-ROC). Results indicate that L1 Regularization with LR and Chi-Squared with RF perform best for specific datasets. The final model, combining all datasets with Forward Selection and RF, achieves high accuracy (91%), precision (87%), recall (97%), f1-score (91%), and AUC-ROC (94%). This study concludes that advanced text-based feature selection and SHAP interpretability significantly enhance CFR model accuracy and transparency, aiding clinical decision-making. Future research should incorporate more diverse datasets, explore advanced NLP techniques, and validate models in various clinical settings to enhance robustness and applicability.

***Keywords*— Cardiac failure recognition; clinical narratives; predictive modelling; SHapley Additive exPlanations (SHAP).**

Manuscript received 7 Dec. 2023; revised 19 Jun. 2024; accepted 18 Oct. 2024. Date of publication 31 Dec. 2024.

I. INTRODUCTION

Cardiac failure recognition (CFR) within clinical narratives stands as a critical challenge in healthcare, demanding accurate and timely identification for optimal patient care. The complexities of unstructured medical data, including diverse linguistic patterns, limited dataset sizes, and underutilized vital sign information, have hindered the effectiveness of existing methodologies in CFR [1]. To address these challenges, this study explores advanced text-based feature selection techniques and the SHapley Additive exPlanations (SHAP) interpretability method as promising approaches.

Clinical narratives, with their nuanced and multifaceted unstructured data, require advanced feature selection to distill crucial information effectively [2]. By applying natural language processing (NLP) strategies, this research transforms unstructured clinical texts, such as patient histories and admission notes, into meaningful numerical representations. This approach addresses issues of language diversity and feature space sparsity, aiming to extract vital clinical insights [3].

In parallel with feature selection, model interpretability is crucial in ensuring that machine learning (ML) predictions are not only accurate but also comprehensible and transparent for clinical practitioners [3]. The integration of interpretability into predictive modeling enhances decision-making in healthcare settings, where understanding the factors behind predictions is essential [4]. This study proposes to combine NLP strategies with SHAP to refine algorithms, optimize feature selection, and enhance model interpretability, thereby improving CFR's accuracy and transparency in real-world clinical applications.

Existing studies on CFR in clinical narratives, while valuable, face significant challenges that limit their applicability in clinical practice. A major limitation is the use of small datasets, which restricts the ability to capture diverse patterns and essential relationships, thereby affecting the accuracy and interpretability of CFR models [4], [5]. Another challenge is the feature space sparsity, which complicates the identification of relevant information within unstructured data [6]. Additionally, the underutilization of vital sign numeric values in clinical narratives limits the predictive accuracy of CFR models, as highlighted in several studies [1], [3], [4], [5]. These limitations underscore the need for innovative approaches that enhance feature selection, model generalization, and interpretability.

This study employs four public datasets from repositories like Kaggle and the University of California, Irvine (UCI). It focuses on developing a decision model for CFR through advanced text-based feature selection methods, analyzing unstructured clinical notes, and addressing feature space sparsity in small datasets. The feature selection techniques used include Chi-Squared (filter method), L1 Regularization (embedded method), and Forward Selection (wrapper method). Emphasis is placed on refining predictive models for CFR without developing new diagnostic tools or interventions.

This study aims to develop a model using NLP strategies tailored for text-based feature selection in CFR. Also, this study deploys advanced ML models trained on features derived from filter, wrapper, and embedded methods to predict cardiac failure within clinical narratives. Further, this study assesses the models' performance on diverse clinical datasets using metrics like accuracy, precision, recall, F1-score, and AUC-ROC to make sure they are applicable in medical settings. Based on the objectives, this study is subject to proposes three research questions, as follows:

- RQ 1: How can text-based feature selection methodologies enhance the accuracy of pattern recognition in clinical narratives for effective CFR?
- RQ 2: What novel ML strategies can address the challenges of sparse clinical datasets and underutilized vital sign numeric values for more robust CFR?
- RQ 3: How can integrating SHAP interpretability techniques in ML models contribute to improving clinical decision-making in CFR?

The paper is formatted as follows: A thorough assessment of the literature including the proposed decision model, which combines text-based feature selection with the SHAP method for CFR, is introduced in Section II. In Section III, the outcomes of the proposed model are shown and discussed. The study is finally concluded in Section IV.

II. Materials and Method

A. Background Study

1) Cardiac Failure Recognition:

Various studies have delved into the intricacies of recognizing and managing cardiac failure through clinical narratives, shedding light on diverse aspects of this condition. Thanh-Dung Le et al. engineered a ML algorithm, leveraging NLP on physician notes to discern cardiac failure from a healthy state [1]. Rachel Johnson-Koenke et al. took a narrative inquiry approach, investigating the psychosocial adjustments of veterans grappling with heart failure, scrutinizing alterations in self-schema, world schema, and meaning [7]. In a pediatric context, Jonathan N. Johnson and David J. Driscoll underscored the pivotal role of clinical history and physical examination in evaluating children with heart failure [8]. Additionally, Carlos Sampaio et al. delved into the caregiving experience of family caregivers of heart failure patients, advocating for their active participation in healthcare planning and execution [9]. These studies collectively contribute valuable insights into the multifaceted landscape of CFR and management through clinical narratives.

The identification of cardiac failure hinges on various approaches, each offering distinct pathways for early detection. Notable methodologies encompass ML applications for prompt identification, quantifying cytoskeleton-associated protein 4 (CKAP4) concentrations in blood samples, and analyzing physiological signals to detect deteriorating cardiac conditions [10]. ML techniques like decision trees, support vector machines (SVM), random forest (RF), logistic regression (LR), Naive Bayes and K-nearest neighbor demonstrate promise in detecting early-stage heart failure [11]. Furthermore, sensor circuits coupled with signal processors can derive signal metrics, while risk stratified circuits yield indications of cardiac risk escalation, collectively offering robust avenues for timely diagnosis.

Fundamentally, heart failure presents as a clinical illness when the heart is unable to pump enough blood to meet the body's needs. The symptoms include decreased cardiac output from either diastolic or systolic dysfunction, or a combination of the two, resulting in a smaller stroke volume [12]. Systolic dysfunction arises from a loss of intrinsic inotropy or viable muscle, while diastolic dysfunction signifies decreased ventricular compliance, hindering proper filling [13]. These dysfunctions elevate ventricular end-diastolic pressure, fostering compensatory mechanisms such as the Frank-Starling mechanism to bolster stroke volume [14]. Moreover, heart failure symptoms extend to vasoconstriction, heightened systemic resistance, and escalated left ventricular filling pressure according to [14].

2) Clinical Narratives:

Clinical narratives serve as a rich resource for enhancing the detection of cardiac failure, offering profound insights into the intricate impact of this condition on patients' lives and the perspectives of caregivers and medical professionals [1], [6]. Leveraging NLP techniques such as word representation learning and ML classifiers enables the analysis of these narratives, facilitating the identification of cardiac failure within the context of unstructured medical texts [15].

Notably, approaches like the KTI-RNN model, amalgamating keyword extraction and topic modeling, exhibit promising outcomes in accurately discerning heart failure from these unstructured narratives [16]. But there are drawbacks when it comes to using clinical narratives, including informal vocabulary and sparse information in unstructured data, which may compromise classification model accuracy [9]. Moreover, divergent perspectives on heart failure across patients, caregivers, and medical specialists underscore the necessity of integrating multiple viewpoints in the healthcare pathway.

In the domain of CFR using clinical narratives, diverse studies offer multifaceted insights. For instance, one study engineered a ML algorithm rooted in NLP to automatically distinguish between patients with cardiac failure and those in a healthy state based on physician notes [1]. In parallel, a systematic review delved into the prevalence and underlying causes of misdiagnoses in heart failure cases, revealing a frequent misdiagnosis trend, with chronic obstructive pulmonary disease emerging as the primary misdiagnosis [17], [18]. This study's implications underscore the imperative need for accuracy in diagnosing cardiac failure, considering the prevalence of incorrect diagnoses and their associated conditions.

Furthermore, a qualitative analysis conducted on interviews with women who had failed on contraceptives revealed factors such as health literacy, beliefs, interpersonal dynamics, and structural barriers that influence these failures. These findings bear implications for refining clinical discussions surrounding contraceptives [19]. The exploration of clinical narratives elucidates the multifaceted nature of CFR, unveiling the nuanced interplay between patient experiences, medical assessments, and misdiagnosis trends. This spectrum of insights underscores the imperative of leveraging diverse perspectives and refining diagnostic approaches for optimal patient care.

3) Text-Based Feature Selection:

Text-based feature selection is a fundamental step in both NLP and ML. Its primary aim is to simplify vast amounts of text data by identifying and extracting the most relevant features. There are diverse methods used for this purpose, ranging from techniques based on information theory to those rooted in term frequency distribution and evolutionary algorithms. It has been demonstrated that these techniques greatly increase the accuracy of text-based classification tasks such as sentiment analysis and email spam filtering. In order to guarantee uniformity across different NLP applications, experts emphasize the significance of developing standardized techniques for feature selection. Essentially, text-based feature selection greatly enhances the performance of NLP and ML models by pinpointing the most informative aspects within textual data [20], [21], [22].

At its heart, text-based feature selection involves pulling out essential information or characteristics from unstructured text, aiming to contribute significantly to the identification of cardiac failure. This process uses the capabilities of NLP techniques to convert text into organized numerical representations. The main goal is to select the most influential features, thereby refining model performance by highlighting the most informative elements within textual data.

This technique is particularly vital in identifying cardiac failure as it serves as the mechanism to distill crucial insights from unstructured text. Leveraging sophisticated NLP methodologies, it translates the qualitative nature of language into measurable entities, directing attention towards the most critical features. In healthcare, where clinical narratives and medical reports hold valuable insights, the precision and effectiveness of feature selection significantly impact subsequent analysis and predictive modeling [20], [21], [22].

4) SHAP Technique:

ML has shown promise in forecasting mortality risk in patients with heart failure (HF) brought on by coronary heart disease (CHD) [23]. Despite their effectiveness, many ML models lack interpretability, making it challenging to comprehend the rationale behind their decisions. To address this issue, a new approach leveraging ML and the SHAP method has been developed. This method aims to calculate mortality risk while providing detailed explanations for individual model decisions [24]. By using this method, physicians are able to better comprehend the major variables affecting the model's predictions, which increases their confidence in the model's reliability [25]. Notably, linear models and decision trees have also shown high interpretability in identifying cardiac failure [26]. This interpretability is crucial for fostering trust and comprehension of ML-based predictions in CFR.

In ML, interpretability is the ability to understand the logic underlying the predictions produced by ML models. In medical contexts like CFR, interpreting these decisions is essential for healthcare practitioners. Several studies delve into the challenges and strategies involved in achieving interpretability in ML. One common challenge lies in the methods used to interpret models within analysis workflows. Researchers emphasize the necessity of multi-stage support to effectively integrate interpretability methods. Adilova et al. [27], for example, suggest a method for employing descriptive models to explain the reasoning behind comprehensive interpretable models.

Investigation of the effects of interpretability elements on perceptions and performance emphasizes the need for qualitative study in order to fully understand the consequences of these aspects [28]. Conversely, [29] offer a technique to pinpoint salient characteristics in deep learning models, providing insight into their mechanisms of decision-making. These studies provide insightful information and a range of methods for ML interpretability.

Achieving interpretability in ML models is crucial for enhancing trust and understanding of predictions. The SHAP method emerges as a powerful and widely utilized tool in this domain, grounded in game theory. SHAP helps to enhance trust in model predictions by offering local explanations for ML models, which facilitate a thorough comprehension of the underlying biomedical issues [30]. The versatility of SHAP is evident in its application across various domains, including clinical metabolomics studies. In such studies, SHAP has been effectively employed to explain ML models like partial least square discriminant analysis (PLS-DA) and RFs, shedding light on complex relationships within the data [31].

Moreover, SHAP demonstrates its utility in the context of distributed ML, where ensuring consistency among explanations from different participants is crucial. This application not only improves the overall interpretability of the ML models but also fosters trust in collaborative decision-making processes [32]. By offering insights into feature importance and contributions, SHAP facilitates a more transparent and comprehensible view of ML model outputs, a vital aspect in fields like healthcare, where it is crucial to comprehend predictions precisely. For researchers and practitioners looking to decipher the complexities of complicated models and promote confidence in their applications, the incorporation of SHAP in ML interpretability thus represents a useful strategy.

5) Text-Based Feature Selection and SHAP Technique:

Text-based feature selection and interpretability play pivotal roles in enhancing CFR within the domain of ML and text mining applications. Various research endeavors have focused on devising methodologies to tackle these challenges. Using feature selectors based on error-bound standards related to SVM is one such method [42]. The purpose of these selectors is to increase the accuracy of heart rate variability (HRV)-based congestive heart failure (CHF) diagnosis [33]. Additionally, genetic algorithms have been explored as an avenue for feature selection, showcasing effectiveness in comprehending complex clinical datasets [26]. Moreover, ML models like artificial neural networks (ANN) have been leveraged to predict heart contraction and diagnose heart disease, exhibiting superior prediction accuracy and holding potential as supportive tools for physicians [34]. These methodologies collectively contribute to advancing accuracy and interpretability in the realm of CFR, offering valuable insights for informed decision-making and patient diagnosis.

Addressing challenges within ML and text mining domains, particularly in aligning text features with image features and managing the diverse nature of text descriptions and images, has been a focus of ongoing research. To mitigate these challenges, diverse methodologies have been proposed. A person text-image matching method that not only improves the interpretability of text features but also utilizes an external attack node to manage the diversity seen in related person images and text [35]. The performance of variable importance as a feature selection method, comparing various ML methods for their ability to identify relevant features was explored. The findings underscored that interpretable methods demonstrated superior performance in feature selection [35]. Similar to this, Munda et al. [36] presented a feature selection method that includes partial supervision, showing advantages in improving the picked feature's stability, relevance, and interpretability. Additionally, a feature extraction and selection model for information retrieval aimed to augment the practicality and interpretative capacities of the mode was proposed [1], [3], [5].

The SHAP technique emerges as a crucial instrument for feature selection in text-based classification tasks, offering a unique amalgamation of association analysis and data mining. It specifically tackles the problem of redundant data inside particular features [37]. In terms of ML models, the SHAP values method, belonging to the class of additive feature attribution values, plays a crucial role in identifying relevant features. This use goes beyond simple explanation to actively contribute to prediction model building as a useful feature selection process, especially for data center operations optimization [38] and improving the ML outputs' interpretability [39].

In various studies, the SHAP-assisted method has demonstrated its efficacy in selecting features for predictive models. Notably, it outperforms alternative methods, showcasing lower error rates and reasonable execution times [40]. This robustness positions the SHAP technique as a valuable asset in the toolkit for researchers and practitioners seeking to enhance feature selection in ML applications, particularly in the context of text-based classification tasks. The method's versatility extends to explaining model outputs, underscoring its significance in promoting transparency and interpretability in the increasingly complex landscape of ML models.

B. Related Works

In leveraging ML algorithms rooted in NLP, the focus lies on detecting cardiac failure within clinical narratives. This involves utilizing physician notes and clinical language data for binary classification, distinguishing patients with cardiac failure from those in a healthy condition. Various word representation techniques, including term frequency-inverse document frequency (TF-IDF), bag-of-words, and neural word embeddings, are employed to enhance analysis. Supervised binary classification algorithms like Gaussian Naive-Bayes, LR, and multilayer perceptron neural networks are applied for training classifiers. Notably, the combination of TF-IDF and multilayer perceptron neural network consistently outperforms other configurations, demonstrating high accuracy, precision, recall, and F1 score. The success of these ML approaches is exemplified in a single French institution, showcasing potential applicability across languages and institutions [1], [5].

Regarding the classification of clinical narratives, a small-scale application-specific compact Switch Transformer model was introduced. When compared to pre-trained BERT-based models such as DistillBERT, CamemBERT, FlauBERT, and FrALBERT, this model performed better after being trained from scratch on a small sample of French clinical text information. The Switch Transformer outperformed traditional Transformers with self-attention mechanisms in capturing a variety of patterns by utilizing a combination of expert mechanisms. The multi-layer perceptron neural network proved to be the best-performing model in this investigation, with the suggested model achieving noteworthy results of 87% accuracy, 87% precision, and 85% recall [3].

Contrary to expectations, traditional NLP techniques demonstrated superior performance over transformer models when applied to smaller radiology report datasets with limited training data [3]. Additionally, an innovative vector representation for sentences, termed language-model-based representation, was introduced to enhance sentiment classification in clinical narratives [41]. These findings collectively underscore the potential of tailored models and traditional NLP techniques in addressing challenges posed by limited datasets and varying text complexities within clinical contexts.

Several studies have concentrated on discerning a patient's condition through natural language representation within clinical narratives. A retrospective clinical study by Thanh-Dung Le et al. focused on using clinical natural language to diagnose heart failure in critically ill infants as early as possible. By utilizing ML algorithms and learning representation, they were able to identify heart failure with remarkable success, exhibiting an outstanding classification performance with 89% accuracy, 88% recall, and 89% precision [5]. This work adds to the increasing number of studies highlighting the usefulness of NLP in clinical contexts, especially when it comes to diagnosing pediatric heart failure. The utilization of learning representation and

ML techniques showcases the potential for advanced computational methods to play a pivotal role in early and accurate diagnoses based on clinical narratives, with implications for improving patient outcomes.

Active learning machine techniques have proven instrumental in predicting cardiovascular heart disease, particularly utilizing the UCI repository database. Employing various ML algorithms such as LR, Naïve Bayes (NB), and ensemble models, the study achieved notable accuracies, reaching up to 90% for LR and 89% for NB in the Cleveland dataset, and 85% and 81% in the Hungarian dataset, respectively. Furthermore, the application of the Synthetic Minority Over-Sampling Technique (SMOTE) led to significant improvements, elevating LR and NB accuracies to 92% and 90% for Cleveland, and 85% and 82% for Hungarian datasets. The pinnacle of this research was the proposed stacked ensemble model, showcasing remarkable metrics with 89.66% accuracy and an 89.16% F1-score on the Framingham dataset [4].

These findings underscore the efficacy of active learning machine techniques in enhancing predictive accuracy for cardiovascular heart disease, with a focus on diverse datasets. The application of ensemble models, along with strategic oversampling techniques, exemplifies the versatility and potential of these approaches in optimizing predictive models for real-world scenarios, providing valuable insights for future advancements in cardiovascular disease prediction.

An innovative model designed for the precise identification of heart failure within unstructured clinical notes through deep learning techniques to address the challenges posed by large-scale electronic health record data analysis, the study emphasizes the critical role of accurate heart failure recognition in informing treatment decisions [6]. The KTI-RNN model strategically incorporates the term frequency-inverse word frequency (TF-IWF) model and latent dirichlet allocation (LDA) model for content expansion, alongside the gated-attention-BiRNN (GA-BiRNN) model for classification. The model's evaluation demonstrates promising outcomes, achieving an F1 score of 85.57% and an accuracy rate of 85.59%, underscoring its potential as an effective tool for enhancing heart failure recognition within clinical narratives.

A number of research works have explored the prediction of heart failure with clinical data; one important contribution is the use of a lightweight ML metamodel. With an accuracy of 89.41% and an area under the curve (AUC) of 88.10%, this metamodel—trained on electronic health record (EHR) data—achieved remarkable results. These results highlight the potential value of metamodels in improving predictive power and accuracy when it comes to clinical data-driven heart failure prediction.

In response to limitations identified in previous research relying on the UCI repository and the Cleveland dataset, the current study introduces a novel metamodel that overcomes these constraints. By leveraging a comprehensive dataset from five renowned cardiac datasets, the proposed metamodel enhances generalizability and accuracy, addressing issues related to data availability and the utilization of basic ML models [2]. This novel method creates opportunities for more accurate clinical data-driven heart failure prediction, offering a more complex and useful solution.

C. Discussion

The review of the background study systematically examines diverse research contributions across several key areas. It starts by emphasizing the pivotal role of CFR within clinical narratives, showcasing how studies by Thanh-Dung Le et al. [1], Johnson-Koenke [7], and Johnson and Driscoll [8] offer multifaceted perspectives on the complexities of this condition. These contributions range from ML algorithms based on NLP to understand patient notes, psychosocial modifications of heart failure patients, to the significance of clinical history and physical examination in evaluating children with heart failure. These studies collectively highlight the diversity of approaches and insights into recognizing and managing cardiac failure through clinical narratives.

The in-depth study then focuses on critical elements such as clinical narratives, text-based feature selection, SHAP technique, and their interplay in enhancing CFR. It highlights how clinical narratives serve as rich sources of information and how leveraging NLP techniques aids in discerning cardiac failure within unstructured medical texts. The significance of text-based feature selection in distilling crucial insights from these narratives is emphasized, along with its role in refining predictive models for cardiac failure detection. Moreover, the review delves into the crucial aspect of ML interpretability, highlighting SHAP method aimed at elucidating the rationale behind model predictions in CFR. Collectively, these contributions underscore the importance of accuracy, interpretability, and the integration of diverse perspectives in refining diagnostic approaches, impacting decision-making and patient care in the realm of CFR within clinical narratives.

The analysis in related works explores the complexities inherent in studies associated with CFR, encompassing the challenges of underutilization of vital sign numeric values, limited dataset sizes, and feature space sparsity. It delves into the significance of text-based feature selection to distill crucial information from unstructured data sources, addressing vital sign data underutilization and navigating through small datasets effectively. Moreover, the investigation emphasizes the pivotal role of interpretability within ML models, aligning accuracy with transparency for practical applicability in healthcare settings. This review highlights how well the work aligns with previous research and provides new understandings of feature representation, ML classification, and the effect of feature selection on CFR. One noteworthy approach addresses clinical text classification problems in a way that is consistent with the efficiency goals of the study. Successful ML algorithms for heart failure recognition and strategies for improving detection reinforce the proposed approach. The focus on unstructured data analysis, feature selection, and heart failure recognition in related works provides a connected backdrop to the study, collectively contributing to shaping robust methodologies. Insights into ML modeling and prediction accuracy support key aspects of the proposed study, forming a comprehensive foundation for enhancing CFR in clinical narratives.

The proposed research aligns seamlessly with this backdrop, aiming to address critical limitations identified in the existing studies. It seeks to contribute by developing a robust model utilizing NLP strategies for text-based feature selection, enhancing accuracy, and integration of the SHAP technique for interpretability of cardiac failure detection. The goal of this study is to close gaps in vital sign numeric values underutilization, feature space sparsity, and dataset sizes by integrating advanced ML models and assessing their effectiveness using a variety of clinical datasets.

This section outlines the comprehensive methodology designed to enhance pattern recognition in clinical narrative datasets for effective CFR. The approach integrates advanced text-based feature selection techniques with SHAP interpretability to improve model accuracy and provide actionable insights. The methodology begins with an extensive literature review to contextualize the research, followed by detailed data acquisition and exploration. Separate models are developed for each dataset using ML classifiers and feature selection methods, culminating in a final, robust model that combines all datasets. Each methodological step is carefully detailed to ensure clarity and reproducibility.

D. Data Acquisition and Exploration

To achieve the objectives of this research, four datasets were carefully selected based on their relevance to CFR and their diverse attributes. Each dataset played a critical role in training and validating the models developed in this study. The "Framingham" dataset was used to forecast the 10-year risk of developing coronary heart disease (CHD). It was derived from an ongoing cardiovascular study carried out in Framingham, Massachusetts, and made publicly available on Kaggle. With more than 4,240 records and 15 attributes, this dataset provides a solid framework for model building. An additional dataset, the "Heart Disease" dataset from Kaggle, is essential for determining heart disease risk based on clinical data. It consists of 918 records and 11 clinical variables intended to predict heart disease events.

Records from Cleveland, Hungary, Switzerland, and Long Beach V are included in the extensive 1988 compilation known as the Cleveland dataset, which has 1,025 records altogether and 76 attributes. However, a subset of 14 attributes was primarily utilized, as it has been widely referenced in published experiments. Additionally, the Heart Failure Clinical Records dataset, which focuses on predicting mortality due to heart failure, contains 12 features and 313 records, highlighting the significance of early detection and management of cardiovascular disease risk factors.

Initial data exploration involved a thorough analysis of the datasets to ensure their suitability for model development. Descriptive statistics were calculated to summarize the data, while visualizations such as histograms and box plots were employed to assess data distribution. Missing values were identified and addressed using imputation techniques, and correlations between variables were examined through correlation matrices and heatmaps. Class distribution analysis was conducted to detect any class imbalance issues that could impact model performance. Furthermore, data cleaning processes included text normalization, outlier detection and removal using statistical methods like the Z-score method, and handling of missing data. Data transformation involved scaling and normalization of numerical features, in addition to utilizing methods such as label encoding to encode categorical variables in order to prepare the data for efficient model training and assessment.

E. Model Development

The methodology used to develop a robust and interpretable model for CFR from clinical narrative datasets involved multiple experiments, beginning with separate model development for each dataset and culminating in a combined dataset model. The key steps in this process included loading and preprocessing data, initial model training, feature selection, hyperparameter tuning, implementing a voting classifier, and applying SHAP interpretability techniques.

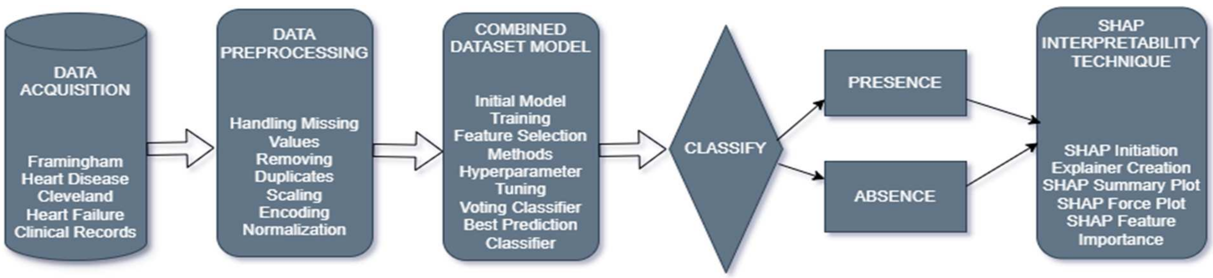

Fig. 1 Conceptual Combined Dataset Model

F. Loading and Preprocessing Data

After loading the four datasets utilized in this study into the environment, missing value handling, categorical variable encoding, numerical feature scaling, and outlier removal were all done as part of the preparation stage. This guaranteed that the datasets were free of errors and appropriate for training models.

G. Initial Model Training

Using three classifiers—SVM, LR, and RF—the first model training was carried out separately on each dataset and on a combined dataset model. Cross-validation scores were used to evaluate these models' performance and create a baseline for comparison between individual and aggregated datasets.

H. Feature Selection Methods

To determine which features were most pertinent for each dataset, three feature selection techniques were used: Forward Selection, L1 Regularization (Lasso), and Chi-Squared Test. By concentrating on the best predictive features, these techniques assisted in enhancing the model's performance by narrowing down the feature set.

I. Hyperparameter Tuning

GridSearchCV was used to optimize the classifiers for each feature selection technique through hyperparameter tuning. By taking this step, the models were optimized for optimal performance.

J. Voting Classifier

To leverage the strengths of different classifiers, a voting classifier was implemented. Both hard and soft voting strategies were explored. This approach aimed to combine the advantages of each classifier, potentially leading to better overall performance.

K. Best Prediction Classifier

By comparing the highest performance metrics of the best single classifier, hard voting classifier, and soft voting classifier for each of the developed models, the best prediction classifier was determined. The final classifier was determined by looking at parameters like accuracy, precision, recall, F1-score, and AUC-ROC and selecting the one that performed better overall. The final model was chosen from among these final classifiers based on its overall performance indicators.

L. Applying SHAP Interpretability Technique

The resulting model was interpreted using SHAP. Insights into how each feature affects the model's predictions are provided by SHAP values, which improve interpretability and transparency. The process involved initializing SHAP, producing SHAP explainers, and producing SHAP summary and force plots.

M. Model Performance Evaluation

To ensure accuracy, reliability, and generalizability, the performance of the ML models built in this study was thoroughly assessed using a range of measures across numerous datasets. Four distinct datasets were used to train the models: the Framingham, Heart Disease, Cleveland, and Heart Failure Clinical Records datasets. Each dataset offered special insights and difficulties that enhanced the final model's resilience. Many assessment criteria, including accuracy, precision, recall, F1-score, and AUC-ROC, were used to evaluate the models' performance. Accuracy, precision, recall, F1-score, and AUC-ROC were used to evaluate the models' performance. Precision evaluated the percentage of genuine positives among the positive forecasts, whereas accuracy evaluated the predictions' total correctness. The model's recall assessed its capacity to identify true positives; its precision and recall were balanced by the F1-score, and its discriminative power across various threshold settings was gauged by AUC-ROC.

The model evaluation process involved training the models on features selected through various feature selection methods, with hyperparameter tuning performed using GridSearchCV. Performance was then measured on test sets, and voting classifiers were utilized to combine the strengths of individual models. The comprehensive evaluation against established benchmarks further ensured the robustness of the models, confirming their accuracy, reliability, and interpretability. The inclusion of diverse datasets validated the models' generalizability across different clinical scenarios, enhancing their potential applicability in real-world healthcare settings.

III. Results and Discussion

The study's results are shown in this section, with particular attention paid to how well the developed models predicted cardiac failure using clinical narrative datasets. It distinguishes the methodology employed from the results obtained, ensuring clarity in the presentation of key outcomes. The impact of various feature selection methods and ML classifiers on model accuracy and interpretability is discussed, with comparisons made to existing literature to underscore the advancements and contributions of this research.

A. Results

The models were evaluated using four diverse clinical datasets: Framingham, Heart Disease from Kaggle, Cleveland Heart Disease, and Heart Failure Clinical Records. Each dataset underwent preprocessing, feature selection, and model training using SVM, LR, and RF classifiers. The best-performing models were identified based on cross-validation scores and further refined through hyperparameter tuning. The final model was developed by combining all datasets and leveraging SHAP interpretability to enhance the understanding of feature importance.

TABLE I
Models Performance Metrics

Dataset	Best Feature Selection Method and Classifier	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Framingham	L1 Regularization with Logistic Regression	0.88	0.50	0.01	0.02	0.73
Heart Disease	Chi-Squared with Random Forest	0.88	0.82	0.78	0.80	0.92
Cleveland	L1 Regularization with Logistic Regression	0.90	0.88	0.93	0.90	0.96
Heart Failure	Chi-Squared with Random Forest	0.90	0.86	0.97	0.91	0.93
Combined	Forward Selection with Random Forest	0.91	0.87	0.97	0.91	0.94

B. Individual Dataset Models

The performance analysis of individual dataset models revealed variations across different datasets and feature selection methods. For instance, the Framingham dataset showed that L1 Regularization combined with Logistic Regression yielded an accuracy of 0.88, while in the Heart Disease dataset from Kaggle, Chi-Squared with Random Forest achieved the most effective results. The Cleveland Heart Disease dataset performed best with L1 Regularization and Logistic Regression, and the Heart Failure Clinical Records dataset saw the highest accuracy with Chi-Squared and Random Forest.

C. Combined Dataset Model

The combined dataset model, using Forward Selection with Random Forest, exhibited superior performance with an accuracy of 0.91. This outcome implies that combining various datasets and using advanced feature selection methods can greatly increase the interpretability and accuracy of the model. The optimal strategy was the combined dataset model, which outperformed the individual dataset models in terms of total performance.

D. Discussion

1) Impact of Text-Based Feature Selection and SHAP Technique:

The study highlights the significant enhancement in accuracy and interpretability achieved through the integration of advanced text-based feature selection methods. Techniques such as Chi-Squared, L1 Regularization, and Forward Selection effectively reduced the dimensionality of the datasets while retaining the most relevant features for cardiac failure prediction. The SHAP interpretability technique provided valuable insights into feature importance, enabling a deeper understanding of the factors influencing model predictions.

2) Comparison with Existing Literature:

The study's findings are in line with earlier research that has shown ML models to be effective in clinical narrative analysis. The methodology proposed in this study not only aligns with but also extends the findings of prior work by integrating SHAP values for enhanced feature importance and model interpretability. This approach addresses common challenges such as feature space sparsity and limited datasets, offering a robust solution for clinical applications.

3) Visualization of Results

The number of true positive, true negative, false positive, and false negative predictions is displayed in the **confusion matrix** for the final model, which shows the classification performance. This matrix highlights the model's high sensitivity and specificity in correctly identifying cardiac failure instances while preserving a low rate of misclassification. The model's practical usefulness in clinical settings is reinforced by the balance between true positives and true negatives in the matrix, which supports the model's effectiveness in differentiating between patients with and without cardiac failure.

With an AUC score of 0.94, the final model's **AUC-ROC curve** further illustrates its ability to discriminate. The model's high AUC value suggests that it can accurately and significantly distinguish between positive and negative cases of heart failure. Plotting the true positive rate against the false positive rate at different threshold settings, the ROC curve indicates how well the model works consistently across thresholds, implying dependable performance in a range of clinical circumstances.

The **precision-recall curve** sheds light on how well the model performs when it comes to correctly predicting positive classes, particularly when datasets are unbalanced. The model shows great precision (87%) and recall (97%), with a high area under the curve. This means that it correctly identifies a high percentage of true positives while keeping the number of false positives low. The model's accuracy and reliability in forecasting cardiac failure are ensured by this delicate balance between precision and recall, which is essential for clinical decision-making and patient care.

A thorough breakdown of how each attribute affects the model's output may be found in the **SHAP summary plot**. The important contributions of features such as ST_Slope, ChestPainType, and ExerciseAngina to the prediction of cardiac failure are illustrated in this plot. This figure improves the interpretability and transparency of the ML model by showing the distribution of SHAP values over all patients, which offers a nuanced knowledge of how specific features affect model predictions.

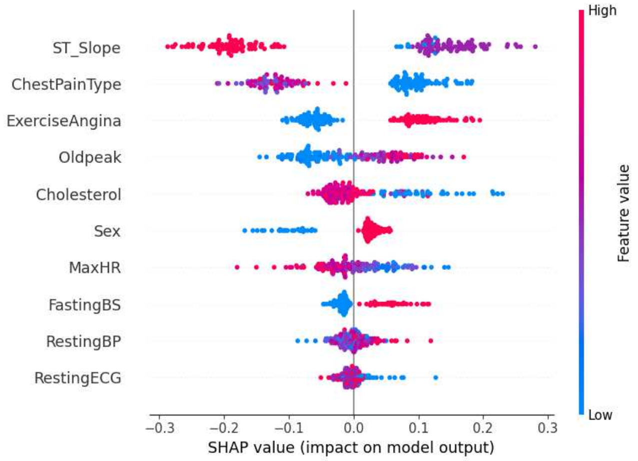

Fig. 2 SHAP Summary Plot

The **SHAP force plot** allows for a detailed understanding of the model's decision-making process by visualizing the contribution of each attribute to the final forecast for particular instances. This representation is especially helpful in clinical contexts, as trusting AI-driven tools requires knowing the logic behind each prediction. For instance, in one prediction, the blue section shows features like ST_Slope, ExerciseAngina, and MaxHR driving the prediction towards the absence of cardiac failure, while the red section highlights feature like Sex, FastingBS, and ChestPainType as factors pushing the prediction towards cardiac failure.

Finally, a clear hierarchy of feature relevance is provided by the **SHAP feature importance plot**, which ranks the features according to their mean absolute SHAP values. The highest significance features, such as ST_Slope, ChestPainType, and ExerciseAngina, are demonstrated to have a significant impact on the model's predictions. This ranking helps determine which clinical factors are most important for predicting cardiac failure, which in turn helps to direct clinical focus in patient assessments as well as model refining.

TABLE II
SHAP FEATURE IMPORTANCE

Feature	Mean Absolute SHAP Value
ST_Slope	0.166517
ChestPainType	0.110089
ExerciseAngina	0.082738
Oldpeak	0.060024
Cholesterol	0.047152
Sex	0.040420
MaxHR	0.038780
FastingBS	0.027112
RestingBP	0.017649
RestingECG	0.012366

By using multiple diverse datasets, the study exhibits resilience and guarantees the reliability and generalizability of its conclusions. Applying SHAP values to feature importance greatly improves the models' interpretability, increasing their transparency and usefulness in clinical contexts. Additionally, the comprehensive evaluation using various performance metrics provides a thorough assessment of the models, ensuring that the results are well-supported and credible.

Despite these strengths, the study does have some limitations. For instance, the Framingham dataset exhibited lower recall and F1 scores, which may indicate challenges related to feature representation or specific characteristics of the dataset. Furthermore, the reliance on historical datasets may limit the applicability of the models to contemporary clinical practices, as medical knowledge and practices evolve over time. It is important to take these limitations into account when evaluating the findings and planning future research.

An extensive investigation of the developed models' ability to predict cardiac failure using clinical narrative datasets is presented in this section. A thorough analysis of the data revealed that the combined dataset model—which incorporated Forward Selection and Random Forest—performed better than the individual dataset models in terms of F1 score, AUC-ROC, accuracy, precision, and recall. The models' improved performance and transparency were due to the incorporation of advanced feature selection methods and SHAP interpretability. The strengths of this study lie in its use of diverse datasets and rigorous evaluation metrics, which ensure the generalizability and reliability of the findings. However, the limitations related to specific dataset characteristics and the use of historical data must be acknowledged.

These findings underscore the importance of integrating diverse datasets and employing advanced feature selection techniques to enhance model interpretability and predictive performance in clinical situations. The knowledge acquired from this research aids in the continuous creation of AI-powered instruments for CFR, which may have consequences for patient care and clinical judgment.

IV. Conclusion

This study successfully developed accurate and interpretable models for CFR within clinical narrative datasets by leveraging advanced text-based feature selection methods and the SHAP interpretability technique. Using four diverse datasets—Framingham Heart Study, Heart Disease from Kaggle, Cleveland Heart Disease, and Heart Failure Clinical Records—this research addressed key challenges in clinical data analysis, such as limited datasets, feature space sparsity, and the underutilization of vital sign relations.

Through rigorous preprocessing, feature selection, and model training, effective combinations of methods and classifiers were identified, enhancing both predictive performance and model transparency. For example, L1 Regularization with Logistic Regression performed best for the Framingham and Cleveland datasets, while Chi-Squared with Random Forest excelled for the Heart Disease and Heart Failure datasets. The final model, which integrated all datasets using Forward Selection with Random Forest, achieved the highest accuracy and interpretability, highlighting the advantages of combining diverse clinical data sources.

The study's findings highlight the significance of combining advanced feature selection methods with SHAP interpretability to create models that are transparent, reliable, and accurate for use in clinical decision-making. By reducing dimensionality while retaining crucial features, techniques like Chi-Squared, L1 Regularization, and Forward Selection proved to be instrumental in enhancing model performance across multiple datasets.

The results of the study have important ramifications for clinical practice and research. The methodology proposed here is robust and generalizable, making it suitable for use in a variety of clinical contexts and eventually improving patient outcomes by increasing cardiac failure prediction. Still, the study also pointed out several directions for additional investigation. To further improve feature extraction and model performance, future research should investigate the integration of more recent and varied clinical datasets as well as advanced NLP techniques like transformer models.

Moreover, expanding the exploration of interpretability techniques beyond SHAP values and developing user-friendly interfaces for model predictions are crucial steps toward fostering trust in AI-driven tools in healthcare. Real-time clinical decision-support systems and cross-institutional validation should be explored to validate the models' resilience and flexibility and guarantee their applicability to various patient populations and healthcare settings.

This study's conclusion emphasizes how crucial it is to combine diverse datasets and apply robust feature selection methods to create precise and understandable CFR models. With the potential to improve clinical decision-making and patient care, this work advances AI applications in healthcare by tackling important obstacles in clinical data processing and offering precise paths for future research.

Acknowledgment

We express our sincere gratitude to the Universiti Putra Malaysia (UPM) Faculty of Computer Science and Information Technology for their kind financial support that enabled this publishing. We also wish to acknowledge the faculty's commitment to fostering an environment conducive to research by providing the necessary resources and infrastructure. This support was crucial to the successful execution and completion of our study, enabling us to make meaningful contributions to the field.

References

[1] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Jouvet, "Machine Learning Based on Natural Language Processing to Detect Cardiac Failure in Clinical Narratives," Dec. 20, 2021, *arXiv*: arXiv:2104.03934.

[2] I. Mahmud, M. M. Kabir, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che, "Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel," *Diagnostics*, vol. 13, no. 15, Art. no. 15, Jul. 2023, doi: 10.3390/diagnostics13152540.

[3] T.-D. Le, P. Jouvet, and R. Noumeir, "Improving Transformer Performance for French Clinical Notes Classification using Mixture of Experts on a Limited Dataset," Mar. 22, 2023, *arXiv*: arXiv:2303.12892.

[4] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. M. M. B, P. Jayagopal, and G. T. Dalu, "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database," *Sci. Rep.*, vol. 13, no. 1, Art. no. 1, Aug. 2023, doi:10.1038/s41598-023-40717-1.

[5] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Jouvet, "Detecting a Patient's Condition From Clinical Narratives Using Natural Language Representation," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 142–149, 2022, doi: 10.1109/ojemb.2022.3209900.

[6] D. Li et al., "KTI-RNN: Recognition of Heart Failure from Clinical Notes," *Tsinghua Sci. and Technol.*, vol. 28, no. 1, pp. 117-130, Feb. 2023, doi: 10.26599/tst.2021.9010093

[7] R. Johnson-Koenke, "Stories of the Heart: Illness Narratives of Veterans Living With Heart Failure," *Fed. Pract.*, no. 39 (5), no. 39 (5), May 2022, doi: 10.12788/fp.0260.

[8] J. N. Johnson and D. J. Driscoll, "Clinical Recognition of Congestive Heart Failure in Children," in *Heart Failure in the Child and Young Adult*, Elsevier, pp. 161–165, 2018, doi: 10.1016/B978-0-12-802393-8.00013-2.

[9] C. Sampaio, I. Renaud, and P. P. Leão, "Illness trajectory in heart failure: narratives of family caregivers," *Rev. Bras. Enferm.*, vol. 72, no. 1, no. 1, Feb. 2019, doi: 10.1590/0034-7167-2018-0645.

[10] Kumar Mohan, Yousuf Fadhil Salim AL-Mamari, and Mohammed Ahmed Mohammed AL-Najadi, "Techniques of Machine Learning for Detecting Heart Failure," *Int. J. Data Inform. Intell. Comput.*, vol. 2, no. 2, Jun. 2023, doi: 10.59461/ijdiic.v2i2.62.

[11] M. Sakoda *et al.*, "Early Detection of Worsening Heart Failure in Patients at Home Using a New Telemonitoring System of Respiratory Stability," *Circ. J.*, vol. 86, no. 7, pp. 1081-1091, Jun. 2022, doi:10.1253/circj.CJ-21-0590.

[12] R. H. G. Schwinger, "Pathophysiology of heart failure," *Cardiovasc. Diagn. Ther.*, vol. 11, no. 1, pp. 263-276, Feb. 2021, doi:10.21037/cdt-20-302.

[13] I. Sopek Merkaš, A. M. Slišković, and N. Lakušić, "Current concept in the diagnosis, treatment and rehabilitation of patients with congestive heart failure," *World J. Cardiol.*, vol. 13, no. 7, pp. 183-203, Jul. 2021, doi: 10.4330/wjc.v13.i7.183.

[14] S. K. Nadar and M. M. Shaikh, "Biomarkers in Routine Heart Failure Clinical Care," *Card. Fail. Rev.*, vol. 5, no. 1, pp. 50-56, Feb. 2019, doi: 10.15420/cfr.2018.27.2.

[15] TRUST participants *et al.*, "The emotional and social burden of heart failure: integrating physicians', patients', and caregivers' perspectives through narrative medicine," *BMC Cardiovasc. Disord.*, vol. 20, no. 1, Dec. 2020, doi: 10.1186/s12872-020-01809-2.

[16] J. Wistrand, "Ailing Hearts and Troubled Minds: An Historical and Narratological Study on Illness Narratives by Physicians with Cardiac Disease," *J. Med. Humanit.*, vol. 43, no. 1, pp. 129-139, Mar. 2022, doi: 10.1007/s10912-020-09610-0.

[17] L. Frohwirth *et al.*, "Understanding Contraceptive Failure: An Analysis of Qualitative Narratives," *Womens Reprod. Health*, vol. 10, no. 2, pp. 280-302, Jul. 2022, doi: 10.1080/23293691.2022.2090304.

[18] C. W. Wong *et al.*, "Misdiagnosis of Heart Failure: A Systematic Review of the Literature," *J. Card. Fail.*, vol. 27, no. 9, pp. 925-933, Sep. 2021, doi: 10.1016/j.cardfail.2021.05.014.

[19] P. Atkinson, "Illness Narratives Revisited: The Failure of Narrative Reductionism," *Sociol. Res. Online*, vol. 14, no. 5, pp. 196-205, Nov. 2009, doi: 10.5153/sro.2030.

[20] J. Misra, "autoNLP: NLP Feature Recommendations for Text Analytics Applications", arXiv preprint arXiv:2002.03056, 2020, https://doi.org/10.48550/arXiv.2002.03056.

[21] X. Tang, Y. Dai, and Y. Xiang, "Feature selection based on feature interactions with application to text categorization," *Expert Syst. Appl.*, vol. 120, pp. 207–216, Apr. 2019, doi: 10.1016/j.eswa.2018.11.018.

[22] K. Thirumoorthy and K. Muneeswaran, "Feature Selection for Text Classification Using Machine Learning Approaches," *Natl. Acad. Sci. Lett.*, vol. 45, no. 1, pp. 51-56, Feb. 2022, doi: 10.1007/s40009-021-01043-0.

[23] K. Wang *et al.*, "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP," *Comput. Biol. Med.*, vol. 137, p. 104813, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104813.

[24] M. Bodini, M. W. Rivolta, and R. Sassi, "Opening the black box: interpretability of machine learning algorithms in electrocardiography," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 379, no. 2212, Oct. 2021, doi: 10.1098/rsta.2020.0253.

[25] T. Räz, "ML Interpretability: Simple Isn't Easy," *Studies in History and Philosophy of Science*, vol. 103, pp. 159-167, Feb. 2024, doi:10.1016/j.shpsa.2023.12.007.

[26] M.-Y. Lee and S.-N. Yu, "Selection of Heart Rate Variability Features for Congestive Heart Failure Recognition Using Support Vector Machine-Based Criteria," in *5th European Conference of the International Federation for Medical and Biological Engineering*, vol. 37, pp. 400–403, 2011, doi: 10.1007/978-3-642-23508-5_104.

[27] L. Adilova, M. Kamp, G. Andrienko, and N. Andrienko, "Re-interpreting Rules Interpretability," In Review, preprint, Apr. 2022. doi: 10.21203/rs.3.rs-1525944/v1.

[28] J. Qu, J. Arguello, and Y. Wang, "Understanding the Cognitive Influences of Interpretability Features on How Users Scrutinize Machine-Predicted Categories," *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pp. 247–257, Mar. 2023, doi: 10.1145/3576840.3578315.

[29] S. Sengupta and M. A. Anastasio, "A Test Statistic Estimation-Based Approach for Establishing Self-Interpretable CNN-Based Binary Classifiers," *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1753–1765, May 2024, doi: 10.1109/tmi.2023.3348699.

[30] O. O. Bifarin, "Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification," *PLOS ONE*, vol. 18, no. 5, May 2023, doi:10.1371/journal.pone.0284315.

[31] A. Bogdanova, A. Imakura, and T. Sakurai, "DC-SHAP Method for Consistent Explainability in Privacy-Preserving Distributed Machine Learning," *Hum.-Centric Intell. Syst.*, vol. 3, no. 3, Jul. 2023, doi:10.1007/s44230-023-00032-4.

[32] A. Salih *et al.*, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems,* Jun. 2024, doi: 10.1002/aisy.202400304.

[33] M. Tuan Le, M. Thanh Vo, N. Tan Pham, and S. V.T Dao, "Predicting heart failure using a wrapper-based feature selection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, pp. 1530, Mar. 2021, doi:10.11591/ijeecs.v21.i3.pp1530-1539.

[34] M. Ashtiyani, S. Navaei Lavasani, A. Asgharzadeh Alvar, and M. R. Deevband, "Heart Rate Variability Classification using Support Vector Machine and Genetic Algorithm," *J. Biomed. Phys. Eng.*, Aug. 2018, doi: 10.31661/jbpe.v0i0.614.

[35] F. Li, H. Zhou, H. Li, Y. Zhang, and Z. Yu, "Person Text-Image Matching via Text-Feature Interpretability Embedding and External Attack Node Implantation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1-14, Oct. 2024, doi:10.1109/tetci.2024.3462817.

[36] T. Des Touches, M. Munda, T. Cornet, P. Gerkens, and T. Hellepute, "Feature selection with prior knowledge improves interpretability of chemometrics models," *Chemom. Intell. Lab. Syst.*, vol. 240, p. 104905, Sep. 2023, doi: 10.1016/j.chemolab.2023.104905.

[37] H. Mamdouh Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft Comput.*, vol. 27, no. 16, pp. 11259-11274, Jun. 2023, doi: 10.1007/s00500-023-08587-x.

[38] Q. Ni, L. Chen, J. Zhu, J. Pang, Z. Wang, and X. Yang, "Prediction and interpretation of gamma pass rate based on SHAP value feature selection," In Review, preprint, May 2023, doi: 10.21203/rs.3.rs-2974857/v1.

[39] H. Liu, X. Shen, X. Tang, and J. Liu, "Day-Ahead Electricity Price Probabilistic Forecasting Based on SHAP Feature Selection and LSTNet Quantile Regression," *Energies*, vol. 16, no. 13, pp. 5152, Jul. 2023, doi: 10.3390/en16135152.

[40] Y. Gebreyesus, D. Dalton, S. Nixon, D. D. Chiara, and M. Chinnici, "Machine Learning for Data Center Optimizations: Feature Selection Using Shapley Additive exPlanation (SHAP)," *Future Internet*, vol. 15, no. 3, pp. 88, Feb. 2023, doi: 10.3390/fi15030088.

[41] A. Chen, Z. Yu, X. Yang, Y. Guo, J. Bian, and Y. Wu, "Contextualized medication information extraction using Transformer-based deep learning architectures," *J. Biomed. Inform.*, vol. 142, p. 104370, Jun. 2023, doi: 10.1016/j.jbi.2023.104370.

[42] A. Javeed, S.S. Rizvi, S. Zhou, R. Riaz, S.U Khan, and S.J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," *Mobile Information Systems*, 2020(1), 8843115, doi:.10.1155/2020/8843115.